Predict the default behaviour of small personal loan applicant

Business Content Overview and Motivation

1. Rising market of small personal loan

In the past, loans are mainly towards to businesses or enterprises. Even if for personal loans, the borrowing amounts are usually large and involved application processing is complex and time consuming. In recent years, with the booming of consumer markets and diversification of customer needs, the demand for small personal loan is growing rapidly.

As a new variant in the lending market, issuing small personal loan has its own challenges. Due to the small borrowing amount, the variations of applications are increasing, but the application processing is expected to be simplified and accelerated. In addition, the criterions for small personal loan has not been mature and credit screen is biased and influenced by human factors.

2. Leveraging technology to capture the rising market

In order to tackle those challenges and take a position in the rising market, our client, a Telecom company, in partnership with commercial banks, is trying to leveraging the technics of data science and machine learning to understand the behaviours of applicants and make a prediction on whether or not the applicant will default. Based on that, a mobile lending platform will be launched, which is simple, fast and automated.

Our task is to build a prototype to make the prediction on the default behaviour of applicant using the internal data of Telecom company. Most of operations including data acquisition, data processing etc. are on their virtual machine.

Data Acquisition and Description

The data comes from the Telecom company's internal data database, which stores the historical and real-time records of their customers/users. We retrieve the data from database by SQL queries and export it the excel files.

There are three tables in total: Phone Call records; Application downloads and Label of Defaults.

Table1: Phone Call records

This tables provides the call records of each applicant.

The column contains the call information as follows:

BIZ_TYPE （Business zone type ： national/international） ; CALL_TYPE （Call type ： calling/called/call transfer） ; ROAM_TYPE （Roam type ： local/within province roam/outside province roam/international roam） ; MSISDN （Phone number） ;OTHER_PARTY （Phone number of other party） ;START_TIME （starting time of call） ;END_TIME （ending time of call） ;CALL_DURATION （call duration） ;VISIT_AREA_CODE （visit area code when call） ;CALLED_HOME_CODE （other party home code） ;CALLED_CODE （other party visiting area code） ;CHARGING_DURATION （charging duration） ;DAY_ID_D （date of call） ;PRVNCE_ID_D （Province ID） .

Each row corresponds to a record for call information; one user (same phone number_MSISDN) may contain multiply records (call records for different other parties)

Table2: Application downloads
This table provides the Application download information of each applicant.
The column contains the Application download information as follows:
MDN （Phone number） ； TYPE_CODE1 （application download type code (parent)） ； TYPE_NAME1 （application download name (parent)） ； TYPE_CODE2 （application download type code(children)） ； TYPE_NAME2 （application download name(children)） ； CNT ； FLUX （Download Flux_B） ； DUR （Download duration_S） ;DATA_DAY （Date of data collection） .

Each row corresponds to a record for application download; one user (same phone number_MSN) may contain multiply download records (different types of downloaded applications_TYPE_CODE1/2)

Table4: Label of Defaults
This table provides the situation of defaults of each applicant (Label data).

The column names are PHONE （Phone number） and RESPONSE （Whether defaults or not） .
Each row corresponds to the situation of defaults: 0 repay; 1 default.

Outline the approach and deliverables

Stage 1 ： Data Wrangling and Feature creation

- Section1: Call records data cleaning and Feature creation
- Section2: Application download data cleaning and Feature creation

Stage 2 ： EDA and Machine Learning

The deliverables will include code, a report and a slide.

Data Wrangling

Section1: Phone Call Records data cleaning and Feature creation

1.1 Dealing with Missing value(NA)

The missing value (NA) counts is shown in figure 1.1. The majority of missing values come from CALLED_CODE（the other party visiting area code） and CHARGING_DURATION（charging duration）.

As similar information of above two variables are also contained in the variables of CALLED_HOME_CODE（other party home code） and CALL_DURATION（call duration), in order to reserve most of data, we decide drop this two columns.

After drop the two columns of CALLED_HOME_CODE（other party home code） and CALL_DURATION（call duration）, we check the missing values again, finding that the missing values are reduced significantly, shown in figure 1.2

BIZ_TYPE	1
CALL_TYPE	21
ROAM_TYPE	21
MSISDN	21
OTHER_PARTY	900
START_TIME	21
END_TIME	21
CALL_DURATION	21
VISIT_AREA_CODE	22
CALLED_HOME_CODE	164
CALLED_CODE	2203446
CHARGING_DURATION	4323
DAY_ID_D	21
PRVNCE_ID_D	21

BIZ_TYPE	1
CALL_TYPE	21
ROAM_TYPE	21
MSISDN	21
OTHER_PARTY	900
START_TIME	21
END_TIME	21
CALL_DURATION	21
VISIT_AREA_CODE	22
CALLED_HOME_CODE	164
DAY_ID_D	21
PRVNCE_ID_D	21

figure 1.1 missing value counts before and after column drops

Then we remove all of rows that contain missing values and all the missing values are cleaned.

	BIZ_TYPE	CALL_TYPE	ROAM_TYPE	MSISDN	OTHER_PARTY	START_TIME
0	1	1.0	1.0	189	17017	2.017013e+13
1	1	2.0	0.0	181	65	2.017010e+13
2	1	2.0	0.0	181	26	2.017010e+13
3	1	1.0	0.0	181	26	2.017010e+13
4	1	1.0	0.0	181		2.017010e+13

	END_TIME	CALL_DURATION	VISIT_AREA_CODE	CALLED_HOME_CODE	DAY_ID_D
0	2.017013e+13	4.0	755.0	755.0	201701.0
1	2.017010e+13	324.0	28.0	852.0	20170103.0
2	2.017010e+13	8.0	28.0	28.0	20170103.0
3	2.017010e+13	16.0	28.0	28.0	20170103.0
4	2.017010e+13	64.0	28.0	28.0	20170103.0

	PRVNCE_ID_D
0	844.0
1	851.0
2	851.0
3	851.0
4	851.0

figure 1.2 cleaned dataframe

1.2 Aggregate data of same user (same phone number)

1.21 Aggregate call duration based on call type

Sum the call durations for each user according to the call types (calling/called/call transfer). Then unstack the total call time of different call types from rows to columns. Obtain the total call duration for different call types (calling/called/call transfer) for each user, shown in figure 1.3.

	Phone	Calling_Duration	Called_Duration	Transfered_Duration
0	133	6177.0	6979.0	0.0
1	133	44299.0	30344.0	0.0
2	133	31929.0	28958.0	0.0
3	133	51966.0	47868.0	0.0
4	133	62933.0	34590.0	2393.0

figure 1.3 call duration for different call types

1.22 Aggregate call duration based on business zone type

Same procedure is taken for business zone type feature: Sum the call duration for each user according to the business zone types (domestic/international). Then unstack the call duration of business zone types from rows to columns. Obtain the total call duration for business zone type (domestic/international), shown in figure 1.4.

	Phone	Domestic_Duration	International_Duration
0	133	13156.0	0.0
1	133	74643.0	0.0
2	133	60887.0	0.0
3	133	99834.0	0.0
4	133	99916.0	0.0

figure 1.4 call duration for different business zone types

Since the business zone type data contain mixture data type, for example, the domestic business type is represented by integer(int) type ‘1’ or string (str) type ’01’, we need to unify the data type first before the aggregation.

1.23 Aggregate roam information, the other party information, visiting area information and called home code information

Create a series of new features based on the data of ‘ROAM_TYPE’, 'OTHER_PARTY', 'VISIT_AREA_CODE' and 'CALLED_HOME_CODE':

1. roam ratio: use the ratio of number of roam calls/number of all (roam + non-roam) calls to represent the roam levels of each user.
2. diversity of the other party: count the distinct calling numbers and called numbers of each user to represent the diversity of user’s contacts.
3. diversity of visiting area: count the distinct visiting places of each user to represent the diversity of user’s visiting area/travel footprint
4. diversity of geography locations of the other party: count the distinct numbers of geography location for each user’s contacts to represent the geo-diversity of user’s contacts.

	Phone	Contact_Diversity	Contact_Area_Diversity	Visiting_Area_Diversity	Roam_Ratio
0	133	51	16.0	1.0	1.000000
1	133	142	15.0	1.0	0.000000
2	133	201	27.0	14.0	0.030841
3	133	138	21.0	5.0	0.034574
4	133	74	16.0	29.0	0.271057

figure 1.5 Generated new features from ROAM_TYPE’, 'OTHER_PARTY', 'VISIT_AREA_CODE' and 'CALLED_HOME_CODE'

The generated new features from phone call record tables are Contact_Diversity, Contact_Area_Diversity,Visiting_Area_Diversity, Roam_Ratio, Calling_Duration， Called_Duration， Transfered_Duration， Domestic_Duration， International_Duration, shown in figure 1.6.

	Phone	Calling_Duration	Called_Duration	Transfered_Duration	Domestic_Duration	International_Duration	Contact_Diversity	Contact_Area_Diversity	Visiting_Area_Diversity	Roam_Ratio
0	133	6177.0	6979.0	0.0	13156.0	0.0	51	16.0	1.0	1.000000
1	133	44299.0	30344.0	0.0	74643.0	0.0	142	15.0	1.0	0.000000
2	133	31929.0	28958.0	0.0	60887.0	0.0	201	27.0	14.0	0.030841
3	133	51966.0	47868.0	0.0	99834.0	0.0	138	21.0	5.0	0.034574
4	133	62933.0	34590.0	2393.0	99916.0	0.0	74	16.0	29.0	0.271057

figure 1.6 Generated new features from phone call record table

Section2: Application downloads cleaning and Feature creation

2.1Dealing with Missing value(NA)

As shown in figure 2.1, the majority of missing values are contained in TYPE_CODE1 ; TYPE_NAME1 ; TYPE_CODE2 ; TYPE_NAME2. Drop all of rows that contain the missing values.

MDN	0
TYPE_CODE1	39719
TYPE_NAME1	39719
TYPE_CODE2	39719
TYPE_NAME2	39719
CNT	2
FLUX	2
DUR	2
DATA_DAY	2

figure 2.1 missing value counts

2.2 Aggregate data by user (same phone number_MDN)

The main information reflected from application downloads table is downloaded application types and flux data. The downloaded application types contain primary types（TYPE_CODE1 and TYPE_NAME1） and secondary types （TYPE_CODE2 and TYPE_NAME2）

2.21 Aggregate the download flux information based on application types

The new features are created by three steps:

1) As the TYPE_CODE and TYPE_NAME are one-to-one mapping; we use TYPE_CODE to represent downloaded application types. There are 11 primary application types and 70 secondary application types. We choose primary application types (TYPE_CODE1).

2) The unit time flux information is represented by download flux divided by time, that is, FLUX/DUR. Remove the rows where download duration is zero but download flux is non zero.

3) Sum the unit time download flux based on different downloaded application types(TYPE_CODE1). Then unstack that from rows to columns. Obtain the total download flux per unit time of each user for different downloaded applications.

	Phone	T1_Flux	T2_Flux	T3_Flux	T4_Flux
0	133	1.726816e+06	15440.749141	0.0	230455.510739
1	133	1.147808e+06	66237.700343	0.0	105503.417526
2	133	1.086924e+05	0.000000	0.0	34141.083333
3	133	1.750563e+06	203530.954762	0.0	649026.330838
4	133	1.175443e+05	54269.867761	0.0	115883.691985

	T5_Flux	T6_Flux	T7_Flux	T8_Flux	T9_Flux
0	77371.930994	94283.856872	4.478961e+06	40196.812878	5.748542e+06
1	29653.656277	0.000000	9.256816e+04	30073.119895	5.329929e+06
2	0.000000	0.000000	7.178710e+05	3083.600000	6.655859e+03
3	86059.426258	49897.327273	3.419041e+06	150966.822546	5.565008e+06
4	42402.455255	0.000000	2.280022e+06	47554.287681	9.661299e+05

	T10_Flux	T11_Flux
0	4064.743972	0.0
1	74.645408	0.0
2	0.000000	0.0
3	22083.573770	0.0
4	0.000000	0.0

figure 2.2 Generated new features from application downloads table

EDA

We performed several types of Explore Data Analysis

1. Look at label distribution via pie plot and histogram
2. Look at feature distributions via box plot and violin plot
3. Look at feature correlations via correlation map
4. Comparison of median number and max between two classes by bar plot
5. Test statistics for median number

Initial Findings

1. Label distribution in pie plot and histogram

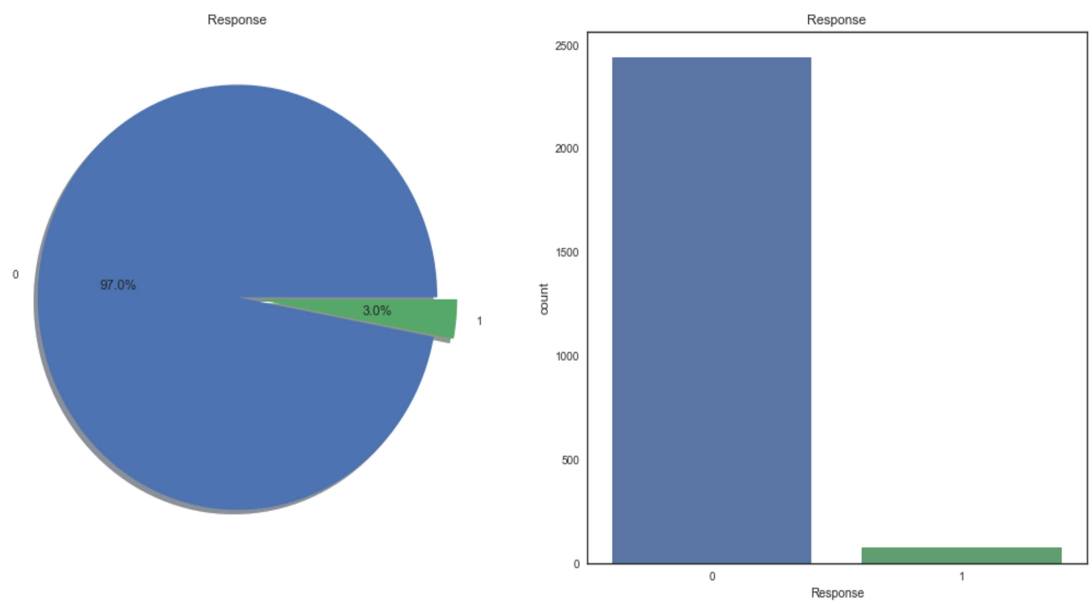


figure 3.1 Response distribution

The distribution of applicant response is given in figure 3.1. It shows that the majority of applicants (97%) are not default.

2. Feature distributions in box plot and violin plot

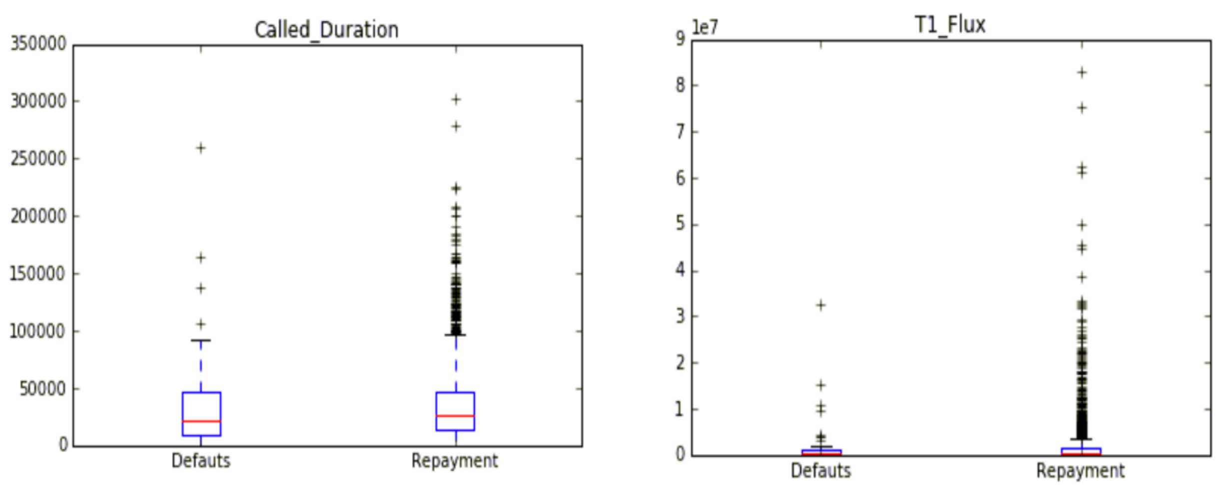


figure 3.2 Example of boxplots for call record feature vs application download feature

The distributions showed from violin maps are similar to that in the box plots. From the boxplots, we find that, compared with call record features, the distribution interval of features in the downloaded application is large, lots of values are treated as outliers.

We think about deal with the data of downloaded application by square root to reduce the distribution interval. The modified feature distributions are shown in figure 3.3. (All of the modified feature distributions are given in the appendix). From the violin maps, it shows that, after the square root, the distribution interval of data is not severely large.

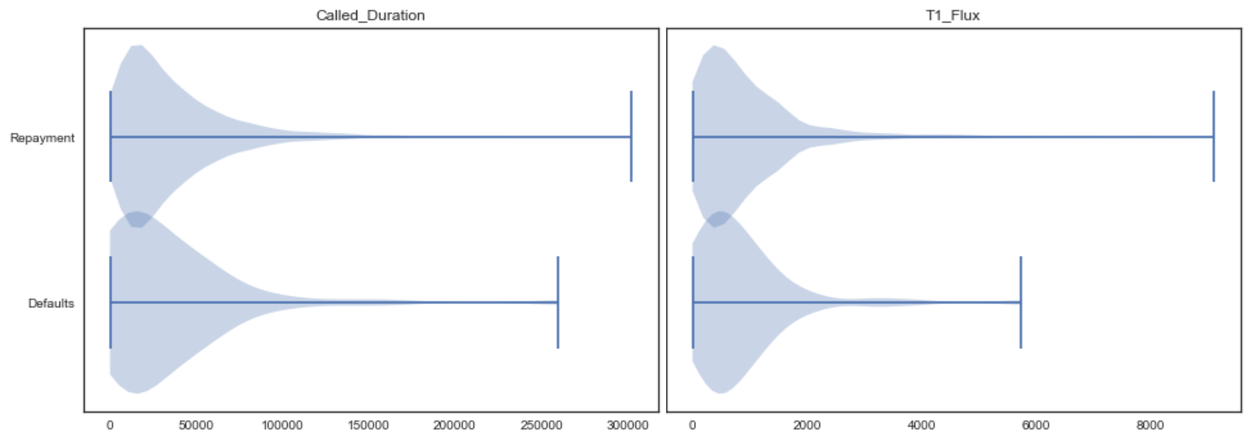


figure 3.3 Example of violin plots for call record feature vs modified application download feature

3. Feature correlations in correlation map

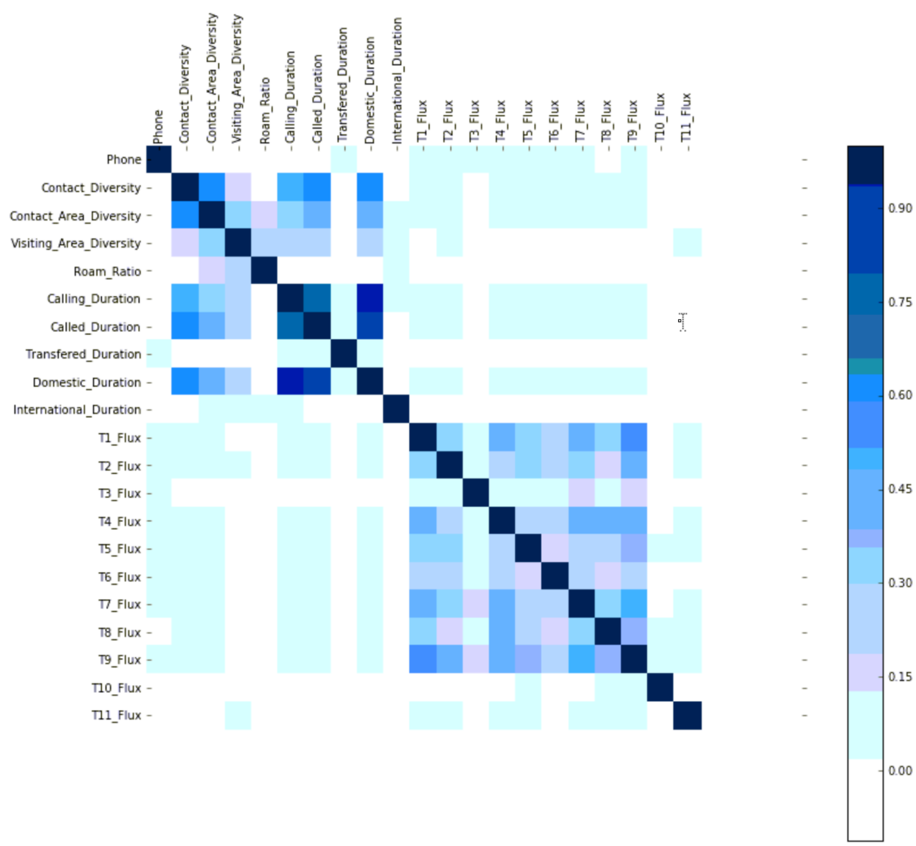


figure 3.4 correlation map of all features

From the correlation map of features (figure 3.2), we can find that:

The colours of upper right section are all light, which means the correlation between the features in Calling records and those in Downloaded application is weak.

The colours of lower right section are relatively light, which means that the correlation between the features among Downloaded application is weak.

The colours of upper left section are relatively high, which means that the correlation between the features among Calling records is strong: such as the number of contacts and the areas diversity of their contacts, the number of contacts the called duration, the number of contacts and domestic call duration, the calling duration and called duration, the calling duration and domestic call duration, the calling duration and called duration.

4. Comparison of median and max number between two classes by bar plot

As the distribution is skewed, we prefer to use median number rather than mean number.

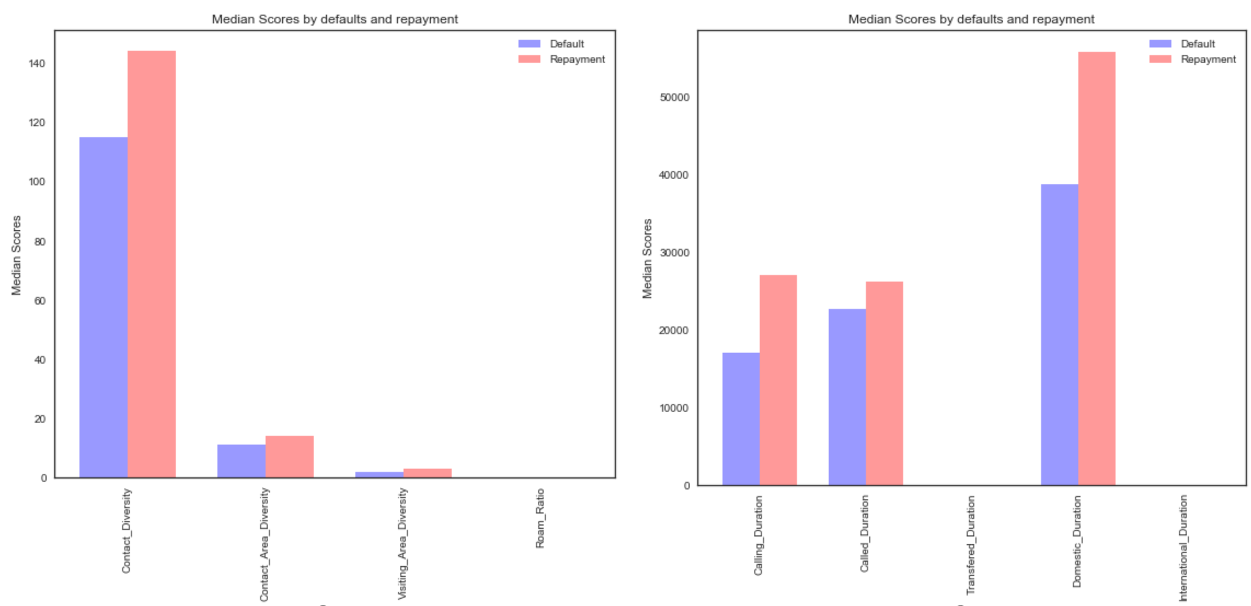


figure 3.5 Comparison of median number of call records features

In figure 3.5, the bar plots of median number of call records features between default class and repayment class shows that compared with default applicants, the repayment applicants have larger average number of contacts and the areas diversity of their contacts are also larger. Their visiting area is more diverse and they have higher roam ratio on average. In terms of call, the repayment applicants have longer average calling, called and domestic call duration.

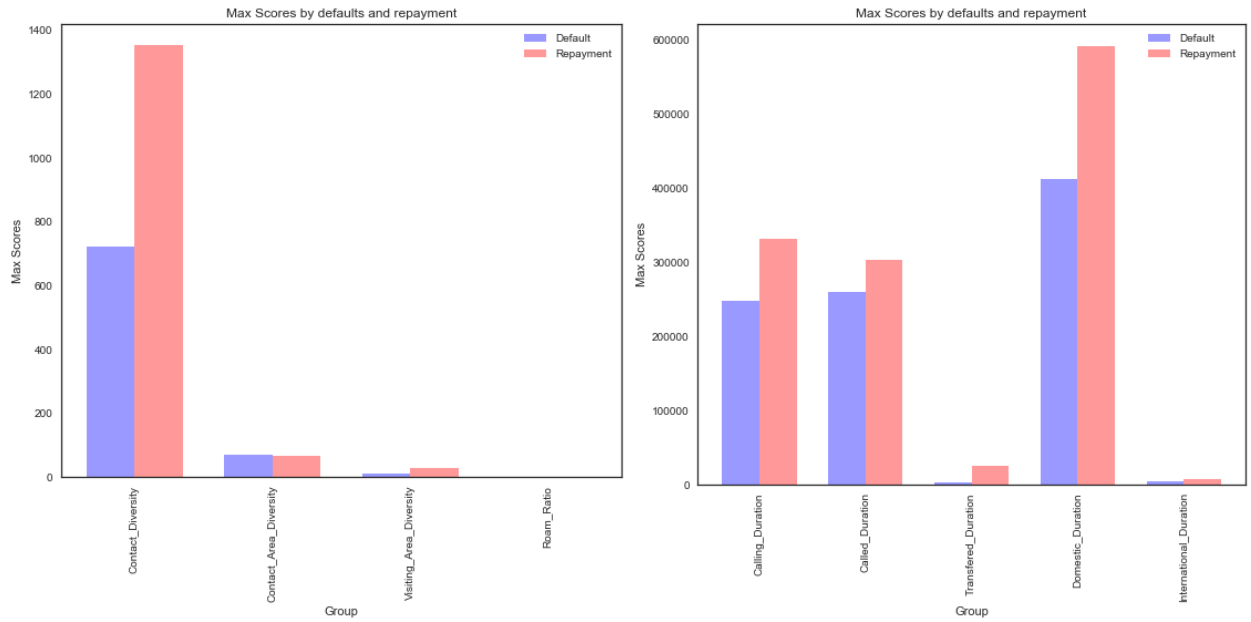


figure 3.6 Comparison of maximum number of call records features

In figure 3.6, the bar plots of max number of call records features between default class and repayment class displays almost the same characteristics with the median number, except for the feature of the areas diversity of applicant contacts, showing a reverse that the repayment applicants have smaller maximum areas diversity of their contacts

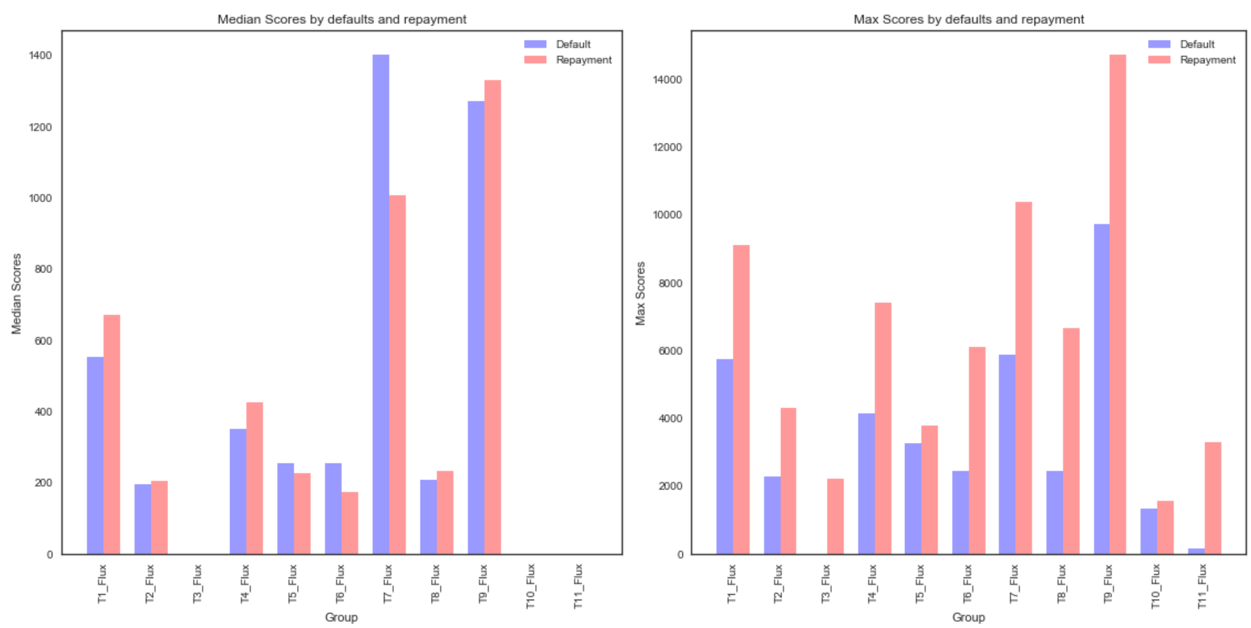


figure 3.7 Comparison of median and maximum number of downloaded application features

The bar plot of media and max number of downloaded application features between default class and repayment class is given in figure 3.7.
On average the repayment applicants have higher downloaded unit time flux in most of categories, excluding in the categories of life service, investment management and news. It is interesting that for default applicants in all of their download categories, the investment management applications cost the highest average download unit time flux while for repayment applicants, the entertainment applications cost the highest.
In the maximum score of downloaded unit time flux, it shows that repayment applicants have higher score in all of categories.

5. Test statistics for the difference of median number between two classes

We do the test statistics for the difference of median number between two classes and the results are given in figure 5.1. For most of features, the difference of median number is not statistically significant and we cannot reject the null hypothesis.

```
Contact_Diversity median-statistic is 1.13229103655 , p-value is 0.14364359189
Contact_Area_Diversity median-statistic is 1.77198825308 , p-value is 0.091568104181
Visiting_Area_Diversity median-statistic is 2.64294541226 , p-value is 0.0520052747668
Roam_Ratio median-statistic is 3.47896815947 , p-value is 0.0310767418796
Calling_Duration median-statistic is 1.36610578377 , p-value is 0.121241046195
Called_Duration median-statistic is 0.489840289186 , p-value is 0.241999278213
Transfered_Duration median-statistic is 0.315891740176 , p-value is 0.287043541121
Domestic_Duration median-statistic is 0.873001692828 , p-value is 0.175062825709
International_Duration median-statistic is 0.117191223386 , p-value is 0.366050514287
T1_Flux median-statistic is 0.489840289186 , p-value is 0.241999278213
T2_Flux median-statistic is 1.22020403498e-05 , p-value is 0.498606440838
T3_Flux median-statistic is 1.61576099144 , p-value is 0.101841935655
T4_Flux median-statistic is 1.969152562 , p-value is 0.0802689207298
T5_Flux median-statistic is 1.22020403498e-05 , p-value is 0.498606440838
T6_Flux median-statistic is 1.98880858476 , p-value is 0.0792327540256
T7_Flux median-statistic is 5.51352658584 , p-value is 0.00943498090158
T8_Flux median-statistic is 0.0533455437921 , p-value is 0.408670397232
T9_Flux median-statistic is 0.489840289186 , p-value is 0.241999278213
T10_Flux median-statistic is 1.64857084968 , p-value is 0.0995767788732
T11_Flux median-statistic is 0.186798578561 , p-value is 0.332797298842
```

figure 5.1 test statistics for the difference of median number between two classes

Machine Learning Model

1. Deploy machine learning models

We deploy Random Forest, Logistic Regression and Gradient boosting on our features and tune the hyperparameters through cross validation.

As the data is imbalanced, the accuracy is not a good metric for evaluation. We choose AUC score. The training results after hyper parameter tuning is given in the following table

Algorithms	**Accuracy**	**AUC score**

Random Forest	0.973880	0.642456
Logistic Regression	0.972748	0.603004
Gradient boosting	0.973880	0.661304

Table 1.1 training results after hyper parameters

The accuracy is high in all algorithms, which is close to the class 0 (repay) proportions. This can be achieved by predict all results as class 0. However, the AUC score, a more reasonable score, is low in all algorithms. In all of algorithms, the Gradient boosting gives best performance score.

It is clear that the imbalance data affect the training error. We will address this issues in the next section.

2. Handle imbalanced data

The strategies to handle the imbalanced data includes undersampling and oversampling(SMOTE). We will also change the cutoff value in the evaluation matrix as a trade-off to achieve the desired target.

Firstly, we prepare the undersampling and oversampling(SMOTE) data and deploy different machine learning algorithms (without hyperparameter tuning) on them. The training results are given in table 2.1

Algorithm	**Without sampling**		**Oversampling**		**Undersampling**	
	Accuracy	AUC score	Accuracy	AUC score	Accuracy	AUC score
Random Forest	0.972181	0.567206	0.972150	0.995681	0.607143	0.624362
Logistic Regression	0.972678	0.608185	0.701841	0.768425	0.598214	0.571429
Gradient boosting	0.972181	0.635041	0.948644	0.990406	0.562500	0.650510
K nearest neighbour	0.972181	0.486633	0.828310	0.917277	0.571429	0.587372

table 2.1 training results on different data samplings

The table shows that after sampling, the balanced data set gives better training results. The oversampling is much better than undersampling. The lower performance of undersampling is caused by small data set after down sampling. The tree methods (Random Forest and Gradient boosting) give very high AUC score.

Then, we use above classifiers on our test data to see the generalization of our models. The test results are shown in table 2.2

Algorithm	**Without sampling**	**Oversampling**	**Undersampling**
	AUC score	AUC score	AUC score
Random Forest	0.52300596853	0.51465002713	0.582528486164
Logistic Regression	0.566142159523	0.592837764514	0.53130765057
Gradient boosting	0.477590884428	0.516115029843	0.525881714596
K nearest neighbour	0.491264243082	0.507650569723	0.447693977211

table 2.2 test results on different data samplings

The prediction performance is just a little bit better than random guess. In general, the oversampling performs better. Without hyperparameter tuning, the logistic regression gives best prediction performance while tree methods suffer from serve overfitting.

The confusion matrix and evaluation metric plot are given as follows

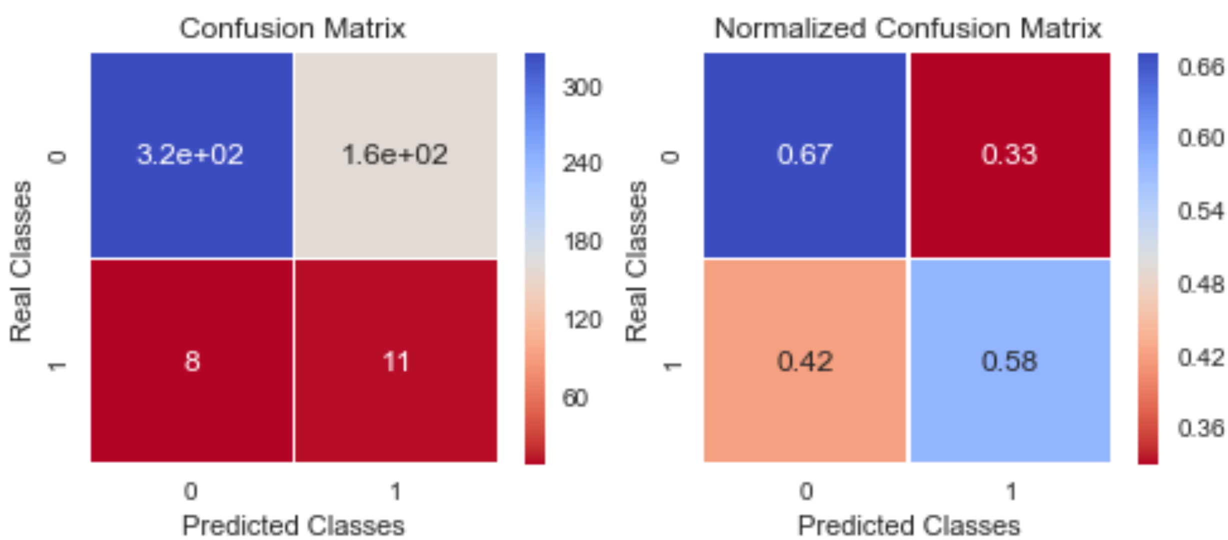


figure 2.1 confusion matrix for logistic regression

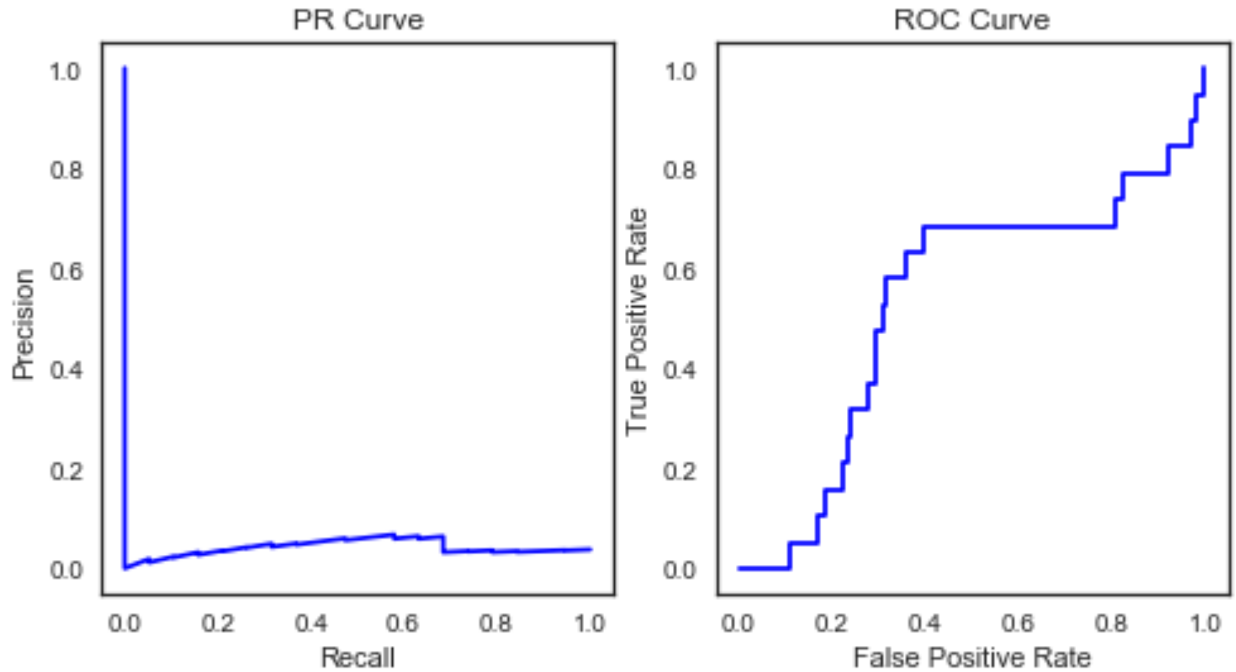


figure 2.2 Precision-Recall and ROC curves for logistic regression

The logistic regression gives a balanced recall score on both of classes and high precision score on class 0 (repay) but low precision score on class 1 (default).

We use grid search to find optimal hyperparameter for logistic regression. The test results after tuning are shown as follows

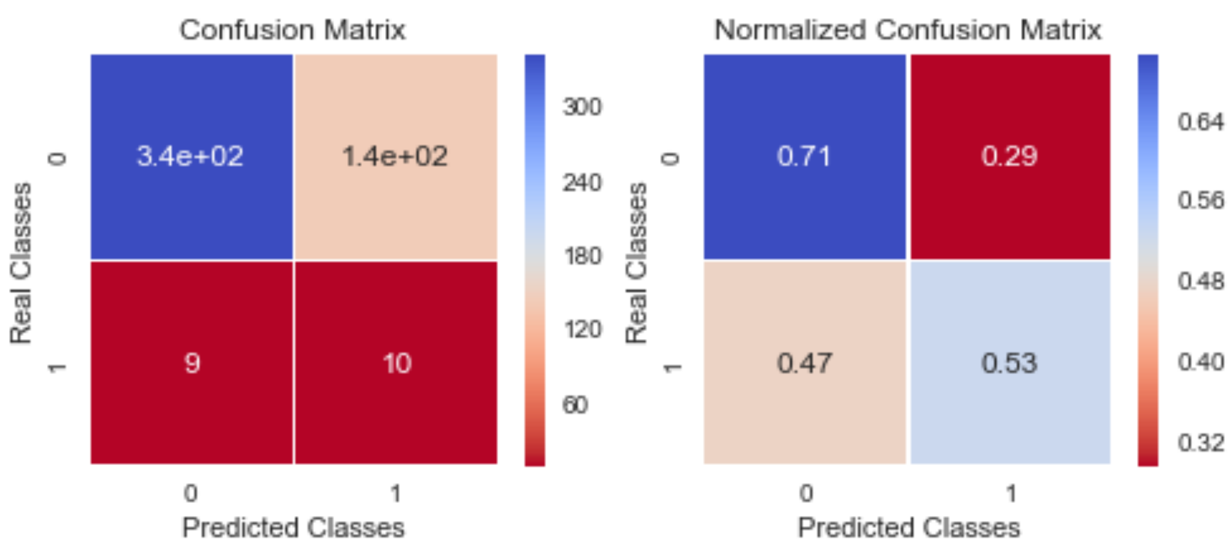


figure 2.3 confusion matrix for logistic regression after hyperparameter tuning

3. Change the Cutoff/Threshold values in evaluation metrics

The sampling can improve the performance but cannot achieve desired target. We can try to change the cutoff/threshold values in evaluation metrics. We show the two extreme of thresholds: when the threshold is >=0.9, the recall metric in the testing dataset is 0 while when the threshold is >=0.1, the recall metric in the testing dataset is 0.89. The confusion metrics and precision recall curve are given as follows

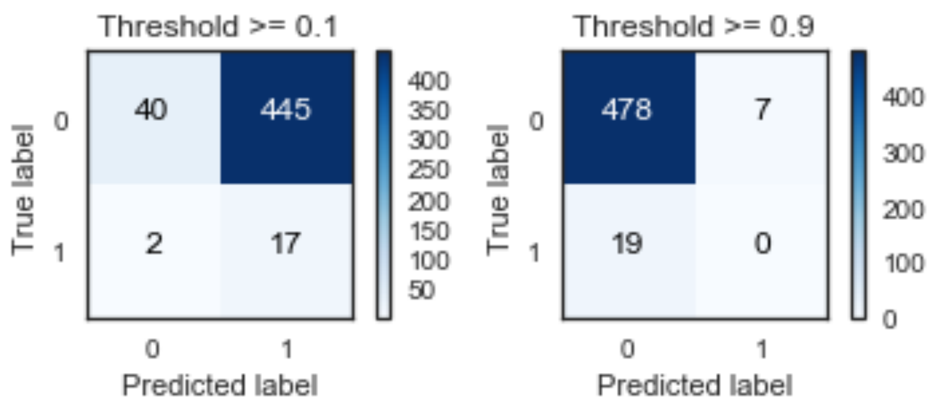


figure 2.4 confusion matrix for logistic regression for different thresholds

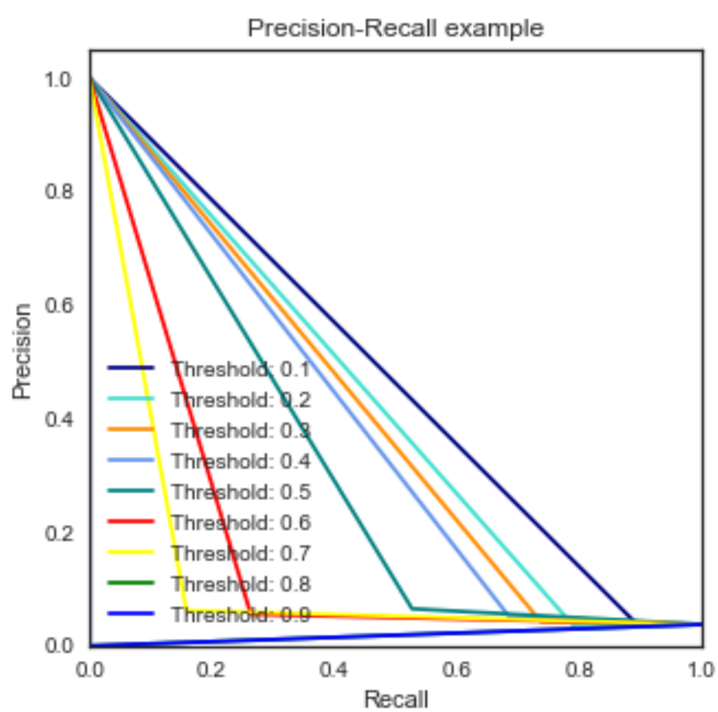


figure 2.5 Precision-Recall curve for logistic regression for different thresholds

4. Improve the performance by Features Processing

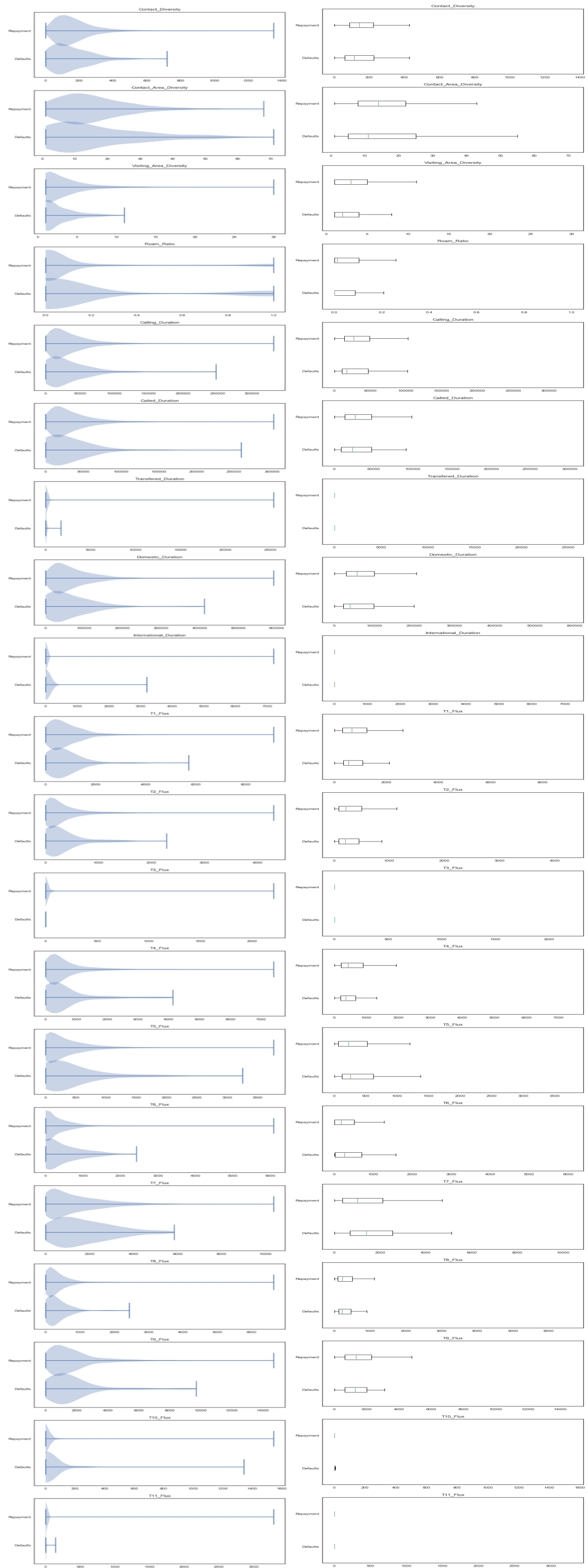
Algorithms	**Features Processing &Samping**	**Test AUC score**
K Nearest Neighbours	`Standarized Features` `Over Sampling`	`0.533516`
Logistic Regression	`Log Transformed Features` `Over Sampling`	`0.571260`
Random Forest	`Log Transformed Features` `Over Sampling`	**`0.647805`**

table 2.3 Test results on different feature processing

5.Future work

A further suggestion is to access more data and create more features.
Also consider try model ensembles (bagging, stack) in future to improve the prediction performance.

Appendix


Contact_Diversity
Contact_Area_Diversity
Visiting_Area_Diversity
Roam_Ratio
Calling_Duration
Called_Duration
Transfered_Duration
Domestic_Duration
International_Duration
T1_Flux
T2_Flux
T3_Flux
T4_Flux
T5_Flux
T6_Flux
T7_Flux
T8_Flux
T9_Flux
T10_Flux
T11_Flux
Repayment
Defaults