Project Report

IS453 - Financial Analytics
Section G1 - Group 4
April 2024

Name	Student ID	Email
CHEAH CHENG PONG	01400523	cpcheah.2021@scis.smu.edu.sg
CHIANG KHENG HE	01428232	khchiang.2021@scis.smu.edu.sg
HO SHOU YIT, KEITH	01391657	keith.ho.2021@scis.smu.edu.sg
JOEL OH KAI SHUN HIROYUKI FUJIKAWA HIROYUKI	01398485	joel.oh.2021@scis.smu.edu.sg
MUHAMMAD FARUQ BIN ABDUL SALIM	01391854	muhammads.2021@scis.smu.edu.sg
RITHVIK BANGALORE SUBRAMANYA	01422126	rithvikbs.2020@scis.smu.edu.sg

Business Context

Through exploratory data analysis in the Bureau dataset, our team identified that the car ownership/owner customer segment formed a majority of the borrowers that the Fintech company can lend to shift to more mainstream borrowers. The objective is to assess the creditworthiness of the identified customer segment, prioritising those with a demonstrated history of financial responsibility and stability through a series of financial indicators.

Initial Exploratory Data Analysis

To merge the 2 datasets, we first have to understand what data is in the respective datasets. For columns which did not have a clear description, we conducted research with the help of ChatGPT and Investopedia.com.

Exploratory data analysis (EDA) was conducted on both datasets separately. Univariate analysis is conducted to understand the possible values within each feature, and the range of the values and identify outliers. One insight found was the invalid outlier in the “DAYS_EMPLOYED” column. Regarding Figure 1 in the appendix, we can see the outlier using a boxplot.

Bivariate analysis was performed next where each variable was analysed with respect to the “TARGET” feature, which is the proportion of good and bad events for each bin in each feature. This allows us to understand the relationship between each independent variable and the TARGET.

One key insight is the unequal number of “SK_ID_CURR” in the application and bureau data. There are 218292 and 307511 “SK_ID_CURR” in the bureau and application dataset respectively. Using this insight, we decided a left merge was suitable for incorporating all bureau data rows that shared matching 'SK_ID_CURR' values with the application data.

Customer Segment Identification

Further EDA is done to identify the customer segment. To achieve this, we would want to have as much data as possible and a segment that can bring in a lot of profits for the Fintech lender. From Figure 2, even though credit card and consumer loans have a high percentage of records, we did not want to target them as these loans are usually not backed by collateral, unlike mortgage and car (auto) loans, making them riskier. Hence, we decided to target customers who want to purchase cars.

From Figure 3, we can also see that 94% of the car loan data records had 0 for the “TARGET” variable, which means most of the customers did not have payment difficulties.

Data Preparation (Pre-Merge)

Before merging the 2 datasets, data preparation was done on both separately. In this step, we handle outliers and duplicate rows. We also removed features that were either irrelevant or had at least 80% missing values in the application dataset. Features that were highly correlated were removed from the application and merged dataset (application + bureau).

Application Dataset

The invalid outlier (“DAYS_EMPOYED”) identified during the EDA was replaced with “Missing” to ensure the raw value does not affect the logistic regression model as the model might interpret the value as a legitimate value.

There were a total of 29 irrelevant features, with 28 being the normalised housing information and the other being gender.

Correlation analysis was performed to identify highly correlated features in the dataset. Feature selection was done to remove highly correlated features and 1 representative feature was kept among the highly correlated pairs to avoid multicollinearity issues when we pass the features into the logistic regression model. The threshold we defined is 80%. An example of a highly correlated pair is “CNT_CHILDREN” and “CNT_FAM_MEMBERS” with a correlation coefficient of 0.879.

The final preprocessing for this dataset is the removal of features with a high number of missing values. The threshold defined is also 80%.

Bureau Dataset

From the initial EDA, we can see each loan can have multiple credit facilities. To obtain a dataset that is suitable for logistic regression, we have to collapse the rows into 1 row, whereby each loan (“SK_ID_CURR”) in the application dataset has only 1 row that contains all the credit facilities information related to the specific loan. This is to ensure there is no duplicated loan application data, which will affect the performance of the logistic regression model.

Feature extraction was done to maintain a unique “SK_ID_CURR” that retains as much information as possible about the loan applicant. The feature extraction uses various methods such as count, mean, minimum, maximum and sum. Referencing

Figure 4 in the appendix, “AMT_CREDIT_SUM_OVERDUE_TOTAL” was created to aggregate the total credit amount overdue across all the credit facilities that an individual has.

We did not remove any columns here with a high number of missing values from this dataset as all the information is crucial for the scorecard generation. In addition, there are underlying reasons why some columns have a lot of missing values. One example is the AMT_ANNUITY column. Regarding Figure 5, we can see that the AMT_ANNUITY is empty because the respective credit facility is closed.

Merged Dataset

Upon preparing and pre-processing both the application and bureau datasets, we merged the application dataset with the bureau dataset using a left join on the “SK_ID_CURR” column, thereby incorporating the additional information from the bureau dataset into the application dataset forming the merged dataset. Feature extraction and feature selection were also done here as there might be features that were highly correlated between the application features and bureau features.

Feature Selection (Merge Dataset)

Since our project focuses on car loans, the “Taken_Car_Loans” feature was created in the bureau dataset preprocessing step to help in this filtering process. Only applicants who have previously taken car loans were kept. After filtering, we have 12168 rows. Within these subsets of car owners, we applied fine class binning:

- 20 bins for numerical,
- individual bins per category for categorical

to find the best predictive strength of the features based on the bin's IV. This fine classing approach helps us determine which variables are most useful based on their bin’s total IV before we proceed with simplifying bins in coarse classing.

After sorting the features based on their IV and judging the top few on the 5Cs, we came up with a preliminary set of 23 features we deemed as most relevant to predicting TARGET (which were in line with ethical lending laws) as seen in Table 1 in the appendix.

Within these 23 fine-classed features, we condensed the number of bins down from 20 to better establish monotonicity, which provides us:

- a clear linear relationship between attribute range and its predictive strength
- simpler scorecard calculation/application process

Since reducing the number of bins hampers the feature’s total IV's slightly, we sought to ensure the values remained within the same predictive range where possible as shown in Figure 6.

Scorecard Creation (Merge Dataset)

Based on IV values and judgement, we selected 14 features for our scorecard. Refer to Table 2 in the appendix.

When looking at their IV values, the top features were:

1. ORGANIZATION_TYPE with IV 0.113
2. DAYS_EMPLOYED with IV 0.1006
3. CREDIT_ACTIVE_Active with IV 0.0905

Based on the barplot in Figure 7, we dropped 9 variables based on IV and judgement. However, we did not drop all the variables with extremely low IVs such as FLAG_OWN_CAR, LANDAREA_AVG, and DAYS_CREDIT_UPDATE_MIN whose IV values are lower than 0.02. This is because these variables add value to our use case despite having low predictive power.

We then carried out WOE encoding on the features for logistic regression and inputted these encodings into a logistic regression classifier to generate the scorecard as shown in Table 3.

We proceeded to use the sc.woebin() library to score and scale the scorecard. We passed in our predetermined target odds, target score and PDO and generated the scorecard which can be seen in the appendix below.

After conducting the coarse classing, we attempted parametric tuning by adding and/or removing variables to come up with the highest accuracy and AUC score which turned out to be with the addition of DAYS_CREDIT_UPDATE_MIN in Figure 8.

The finalised scorecard as shown in Figure 9, had an accuracy and AUC of 0.6225 and 0.6064 respectively.

Scorecard chosen parameters:

TARGET SCORE = 600	**TARGET ODDS =** 1:20 at 600	**PDO =** 20

For every 1 successful applicant, there are 20 rejected (4.66% acceptance rate)
CUTOFF SCORE = 600

Project Limitations and Challenges

Given the skewness of our model as shown in Figure 10, we considered doing downsampling to alleviate the class imbalance. However, given our priority on maintaining as much information as possible, we decided against it.

The merging of application and bureau datasets through the left merge on “SK_ID_CURR” required squashing the data such that each SK_ID_CURR is reflected as only one row. This could have introduced biases, overlooking important information from missing records, which limits the model's training effectiveness.

The limitation of our business domain knowledge in understanding the relationships between certain features and the loan approval result was a challenge. As financial institutions have many different features for consideration, encompassing both quantitative and qualitative types, our approach could have been improved further by engaging with these factors with a stronger understanding of their impacts.

The selection of features for modelling requires both statistical significance and practical relevance. However, our data pre-processing resulted in low IVs (weak predictive power or not useful level) which showed little indication of an individual's creditworthiness concerning car loans. Then, we depended on our limited domain knowledge in the field, resulting in an approximate 60% model accuracy.

Non-financial indicators (categorical variables), such as "OCCUPATION_TYPE" and "ORGANIZATION_TYPE" were hard to measure, which complicated their integration into the model. This meant that the group had to decide which columns to prioritise as inputs for the model, which was equally challenging as it relied heavily on business domain knowledge.

In terms of future works, we can experiment with feature extraction to create features that are medium or highly predictive that would contribute significantly to our model. In addition, we can try different machine learning models to determine if they help in creating a better scorecard. Ensemble methods can be considered as well.

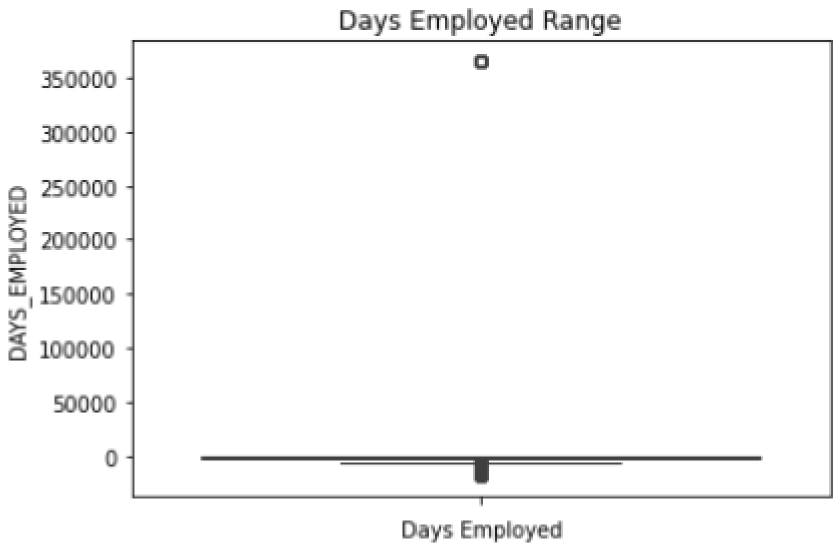
Lessons Learnt

From our analysis, we learnt several key lessons. The importance of EDA and its role in successful decision-making cannot be understated. Business domain knowledge supplements this by correctly interpreting the data and knowing what features to drop. More often than not, we had to research the features for which the definition we were unsure of. The feature selection for the scorecard should balance statistical significance and business relevance while guided by the IV. Initially, we added too many features and it affected the performance.

Appendix

Figure 1

Boxplot of “DAYS_EMPLOYED”



Note. Boxplot showing outliers for “DAYS_EMPLOYED” feature

Figure 2

Pie chart for “CREDIT_TYPE” breakdown

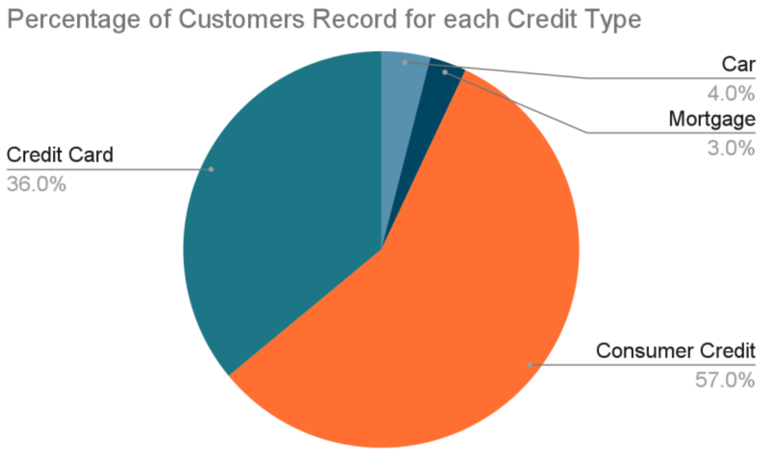


Figure 3

Pie chart for “TARGET” breakdown for Car Loan Data

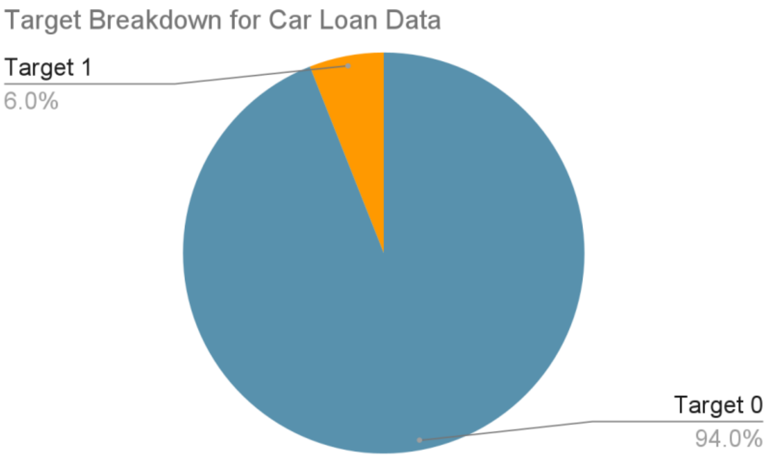
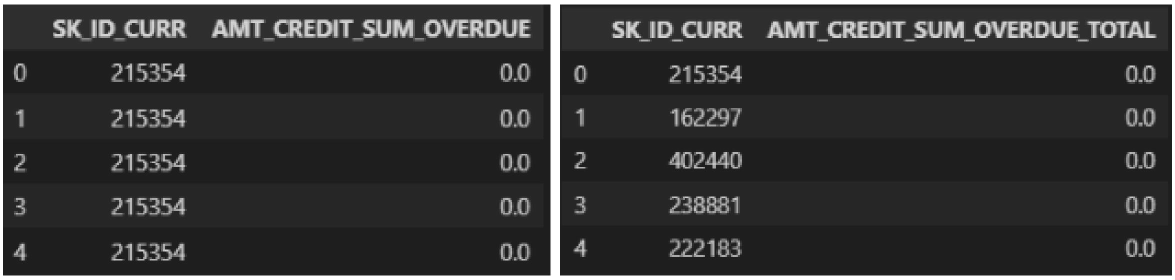


Figure 4

Screenshots of top 5 rows of data frame



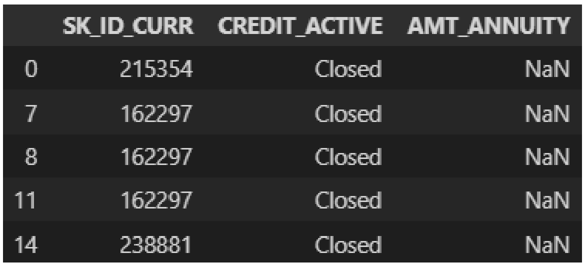
	SK_ID_CURR	AMT_CREDIT_SUM_OVERDUE
0	215354	0.0
1	215354	0.0
2	215354	0.0
3	215354	0.0
4	215354	0.0

	SK_ID_CURR	AMT_CREDIT_SUM_OVERDUE_TOTAL
0	215354	0.0
1	162297	0.0
2	402440	0.0
3	238881	0.0
4	222183	0.0

Note. Data frame before feature extraction (left) and after feature extraction (right)

Figure 5

Top 5 rows where “CREDIT_ACTIVE” is Closed



	SK_ID_CURR	CREDIT_ACTIVE	AMT_ANNUITY
0	215354	Closed	NaN
7	162297	Closed	NaN
8	162297	Closed	NaN
11	162297	Closed	NaN
14	238881	Closed	NaN

Table 1

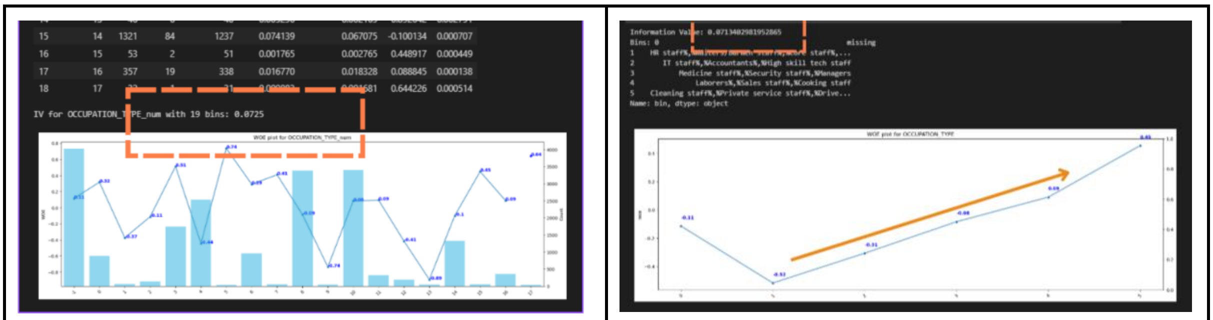
23 chosen features

Character	Capital	Collateral	Capacity	Condition
DAYS_CREDIT_Active_mean NAME_EDUCATION_TYPE CREDIT_OVERDUE_RATIO DAYS_BIRTH OBS_30_CNT_SOCIAL_CIRCLE DEF_30_CNT_SOCIAL_CIRCLE DAYS_CREDIT_UPDATE_MIN AMT_CREDIT_SUM_LIMIT_MEAN NAME_FAMILY_STATUS OWN_CAR_AGE	LANDAREA_AVG FLAG_OWN_REALTY FLAG_OWN_CAR NAME_HOUSING_TYPE	(basing it off Capital since we lack such information)	ORGANIZATION_TYPE CREDIT_ACTIVE_Active OCCUPATION_TYPE CNT_FAM_MEMBERS DEBT_TO_INCOME_RATIO TOTAL_ANNUITY	AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR AMT_REQ_CREDIT_BUREAU_MON

Note. 23 chosen features categorised into 5Cs where minimum IV's is 0.0072 and the highest is 0.13.

Figure 6

Example of coarse classing of fine classes



```
15        14   1321    84     1237   0.074139        0.067075  -0.100134  0.000707
16        15     53     2       51   0.001765        0.002765   0.448917  0.000449
17        16    357    19      338   0.016770        0.018328   0.088845  0.000138
18        17     32     1       31   [illegible]     0.001681   0.644226  0.000514

IV for OCCUPATION_TYPE_num with 19 bins: 0.0725
```

Fine Classing

0.0725 - Weak Predictor

Non - monotonic

```
Information Value: 0.0713402981952865
Bins: 0                                        missing
1    HR staff%,%Waiters/barmen staff%,%Core staff%,...
2      IT staff%,%Accountants%,%High skill tech staff
3         Medicine staff%,%Security staff%,%Managers
4            Laborers%,%Sales staff%,%Cooking staff
5    Cleaning staff%,%Private service staff%,%Drive...
Name: bin, dtype: object
```

Coarse Classing

0.0713 - Weak Predictor maintained

Monotonicity achieved - Clearer linear relationship (ignoring "Missing")

Figure 7

Bar plot of merged data-set features sorted by IV

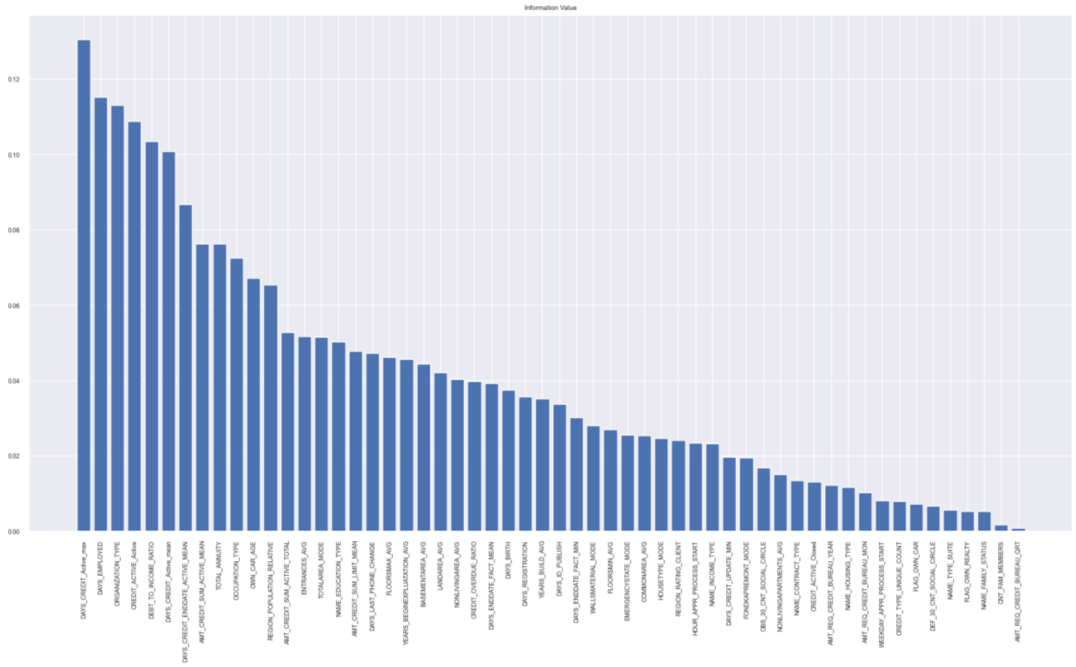


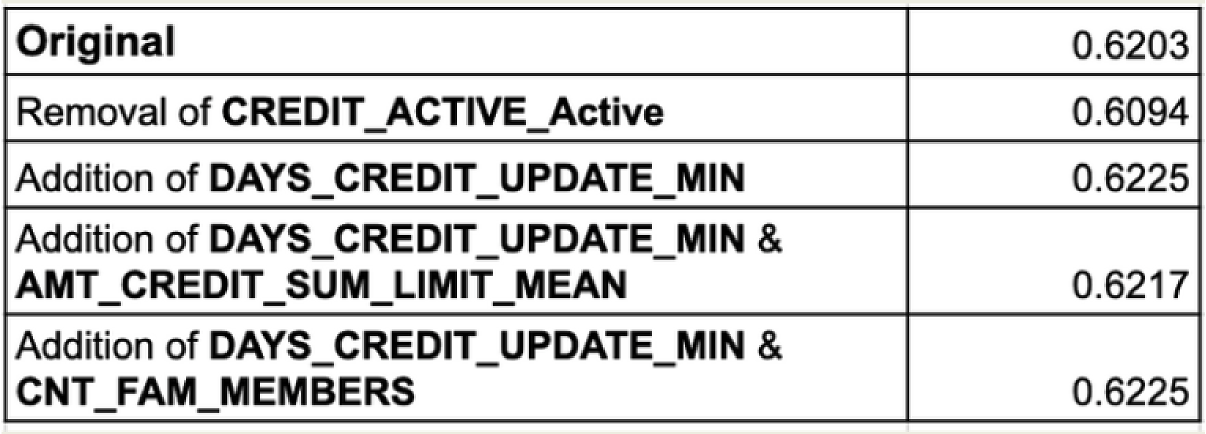
Table 2

Chosen features for scorecard

Features Chosen for Scorecard (IV)
DAYS_CREDIT_Active_mean (0.0869) NAME_EDUCATION_TYPE (0.0495) CREDIT_OVERDUE_RATIO (0.0392) DAYS_BIRTH (0.0272) DAYS_CREDIT_UPDATE_MIN (0.0161) OWN_CAR_AGE (0.056) DAYS_EMPLOYED (0.1006) LANDAREA_AVG (0.0197) FLAG_OWN_CAR (0.0072) ORGANIZATION_TYPE (0.113) CREDIT_ACTIVE_Active (0.0905) DEBT_TO_INCOME_RATIO (0.0764) TOTAL_ANNUITY (0.0348) OCCUPATION_TYPE (0.0713)

Figure 8

Scorecard evaluations during fine-tuning

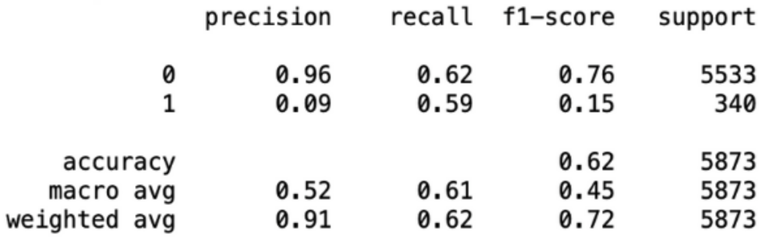


Original	0.6203
Removal of **CREDIT_ACTIVE_Active**	0.6094
Addition of **DAYS_CREDIT_UPDATE_MIN**	0.6225
Addition of **DAYS_CREDIT_UPDATE_MIN** & **AMT_CREDIT_SUM_LIMIT_MEAN**	0.6217
Addition of **DAYS_CREDIT_UPDATE_MIN** & **CNT_FAM_MEMBERS**	0.6225

Note. Addition/removal of features to find the best combination of features for the scorecard.

Figure 9

Performance measures values



```
              precision    recall  f1-score   support

           0       0.96      0.62      0.76      5533
           1       0.09      0.59      0.15       340

    accuracy                           0.62      5873
   macro avg       0.52      0.61      0.45      5873
weighted avg       0.91      0.62      0.72      5873
```

Table 3

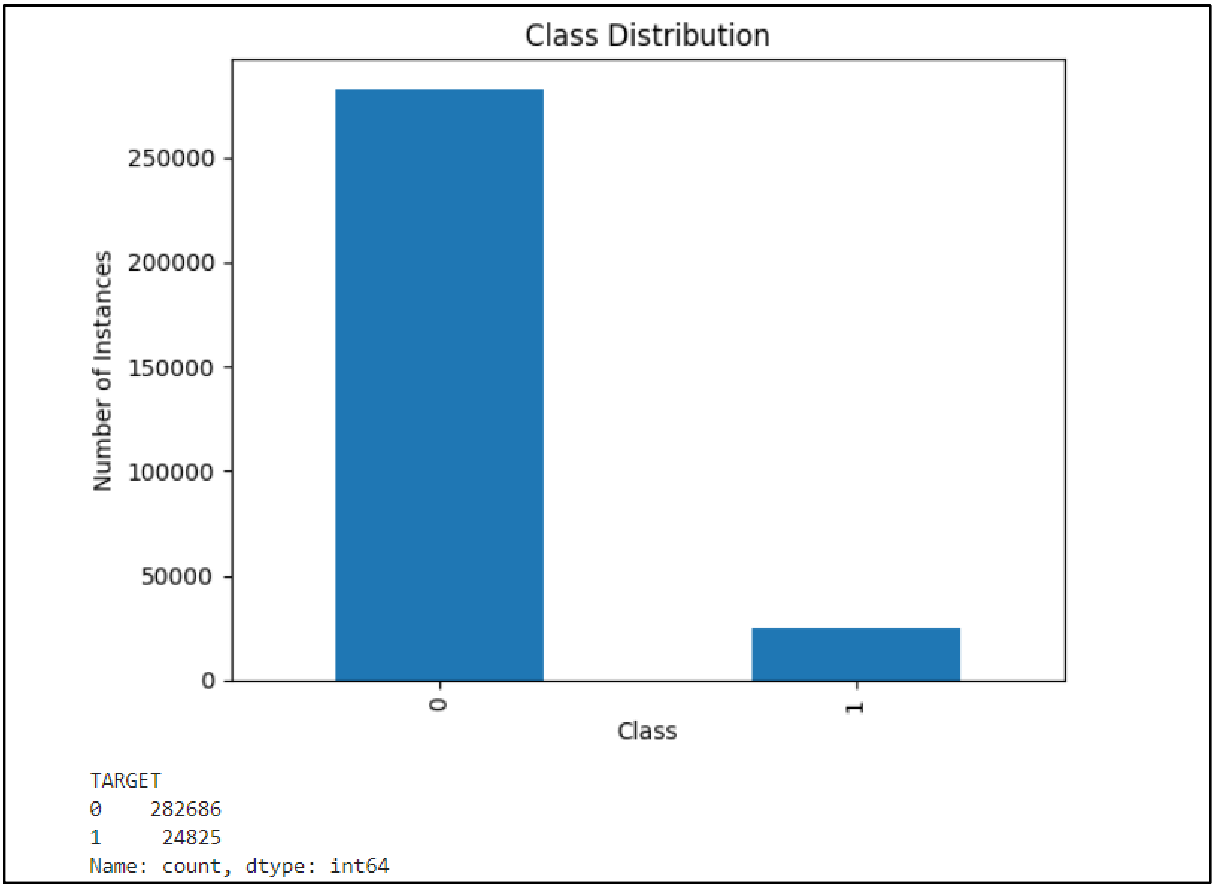
Final Scorecard

Characteristic	Attribute	Score Points
CREDIT_ACTIVE_Active	Missing	55
	< 2.	45
	2.0 - 3.9	38
	4.0 - 4.9	28
	> 4.9	22
CREDIT_OVERDUE_RATIO	< 0.0006	40
	0.0006 - 0.0031	34
	> 0.0031	27
DAYS_BIRTH	< - 21500	42
	-21500 - -15499	38
	-15500 - -13501	37
	> -13501	34
DAYS_CREDIT_Active_mean	Missing	53
	< -1250	53
	-1250 - -699	41
	-700 - -401	32
	> -401	28
DAYS_CREDIT_UPDATE_MIN	< -1100	40
	-1100 - -151	34
	> -151	27

DAYS_EMPLOYED	< -6000	65
	-6000 - -4001	49
	-4000 - -2401	41
	-2400 - -1401	34
	> -1401	31
DEBT_TO_INCOME_RATIO	Missing	40
	< 850000	36
	850000 - 1549999	38
	> 1549999	41
FLAG_OWN_CAR	Y	37
	N	35
LANDAREA_AVG	Missing	35
	< 0.055	38
	0.055 - 0.10	40
	> 0.10	41
NAME_EDUCATION_TYPE	Higher education / Academic degree	43
	Incomplete higher / Secondary	33
OWN_CAR_AGE	Missing	33
	< 6	43
	6 - 7.9	39
	> 8	28
TOTAL_ANNUITY	Missing	24
	< 14000	34
	14000 - 21999	36
	> 21999	39
OCCUPATION_TYPE	Missing	38

	HR staff, Waiters/barmen staff,Core staff,Secretaries	42
	IT staff,Accountants,High skill tech staff	40
	Medicine staff,Security staff,Managers	38
	Laborers,Sales staff,Cooking staff	36
	Cleaning staff,Private service staff,Drivers,Low-skill Laborers,Realty agents	32
ORGANIZATION_TYPE	Hotel ,University ,Religion,Industry	69
	Bank,Emergency,School, Trade	47
	Kindergarten, Business Entity	39
	Medicine, Postal, Construction	34
	Self-employed, Security	28

Figure 10
Distribution of 0 and 1 for “TARGET” variable in application dataset



```
TARGET
0    282686
1     24825
Name: count, dtype: int64
```