Credit Scoring Assignment 1

S1889112

Question 1: PDO

Firstly, we need to know how to go from odds of 1:1 to 64:1 when doubling the odds. To do this, we multiply the odds by 2 repeatedly until we get to 64:1. We want to know 'how many' times to multiply by 2, which is the same as solving for x in $2^x = 64$. The solution to this is $x = \log_2 64 = 6$. In Excel, this is done with the formula **"=log(64, 2)".**

1.1: 20 PDO Scale

We know we must multiply the odds 6 times to get from 1:1 to 64:1. Consider the score at which odds are 1:1 as '**s**'. We know that a 20 point increase doubles the odds, and that we need to double the odds 6 times to get to 64:1. That means $s + 6 * (20) = 580$. Hence solving for **s**, we get a score of 460.

1.2: 15 PDO Scale

Using the same logic as above, we replace 20 with 15 and get $s + 6 * (15) = 580$. This results in **s** = 490.

Question 2: Divergence

The formula for divergence is as follows:

$$Divergence = \frac{(\mu_G - \mu_B)^2}{\frac{1}{2}(\sigma_G^2 + \sigma_B^2)}$$

Where the subscript denotes 'good' or 'bad', μ is the mean, and σ^2 is the variance. The data we have available is:

$$\mu_G = 322.6$$
$$\mu_B = 291.1$$
$$\sigma_G^2 = 820.6$$
$$\sigma_B^2 = 1123.3$$

Using this information, we calculate the divergence as

$$Divergence = \frac{(322.6 - 291.1)^2}{\frac{1}{2}(820.6 + 1123.3)} = \frac{(31.5)^2}{0.5 * (1942.9)} = \frac{992.25}{971.45} = 1.021411$$

Hence we have a divergence of 1.021 (to 3 decimal places). Having quite a low value is intuitively sensible, as the distributions will have substantial crossover given the high variances and the (relatively) close means.

Question 3: Kolmogorov-Smirnov

The general principle of Kolmogorov-Smirnov is to find the maximum distance between cumulative functions. Applying this to the current data means finding the maximum difference between the cumulative proportion of goods and the cumulative proportion of bads. To do this, firstly we order the attributes of the characteristic. We then augment the table with cumulative proportion columns for goods and bads (Using the SUM function, locking the cell equivalent to "<1 year"), and then find the largest difference between these columns (in absolute value). The reason for using absolute value is that we do not care about the direction of the difference, only the magnitude. Augmented columns have their titles highlighted in yellow, and the result is highlighted in green.

Table 1: Kolmogorov-Smirnov Statistic Calculation

Time at Address	% of Goods	% of Bads	Cum % Goods	Cum. % Bads	Difference in cumulative %
<1 year	11.00%	14.50%	11.00%	14.50%	3.50%
1-2 years	3.70%	9.40%	14.70%	23.90%	9.20%
3-5 years	9.60%	15.10%	24.30%	39.00%	14.70%
6-8 years	19.60%	22.80%	43.90%	61.80%	17.90%
9-12 years	7.30%	3.60%	51.20%	65.40%	14.20%
13-24 years	23.40%	17.70%	74.60%	83.10%	8.50%
25+ years	25.40%	16.90%	100.00%	100.00%	0.00%

The largest discrepancy in the cumulative percentage of bads and goods is at a 'time at address' of 6-8 years, with an absolute difference of 17.9%. That is, 17.9% is the KS statistic for the 'Time at Address' characteristic.

Question 4: Factor Importance

In descending order of importance, I would rank the characteristics as 'Credit History', 'Application Score', 'Sector of Employment', and lastly 'Gender'.

As an initial note, what we wish to discern is the combination between the importance of a characteristic in general, and what would be likely to differ between the twins.

Firstly, we consider credit history and the application score from the credit card. We would expect these to vary the most between the twins, as we have not been told any information with regards to purchasing habits, personalities, or other credit-influencing features. Such things are likely to differ between even twins.

Credit history provides information beyond the basic features used in initial scoring for a product. For example, credit history includes how much is spent, what the person spends on, whether they repaid on time, the amount of debt accrued, and regular subscriptions. Given that the twins have the same salary, we can use a credit history as a proxy for the amount of disposable income which is clearly important in their ability to make payments on an additional liability (the car loan). Essentially, although twins would likely have had a similar upbringing, there will be some personality differences which influence purchasing decisions. Even just considering the characteristic, knowing how a person behaves with access to credit should be correlated with their future behaviour on a big-ticket item.

The application score from the credit card can be seen as a measure of risk, just on a different product. As alluded to earlier, a person's behaviour with access to credit should be correlated with their future behaviour (even on a different item). Essentially what this previous score does is estimate the behaviour by considering the risk their behaviour causes. Intuitively, the effect of this can be seen by use of an example. Consider 2 customers; Customer A attained the highest score on a credit card product, and Customer B attained the lowest score. If both customers applied for a new product, we would not expect any sort of (useful) model to score customer B higher than customer A.

With regards to gender and sector, they may both be important characteristics in credit scoring, however we have information on the candidates being twins, and having similar work.

Of the two, I feel **sector** would be more likely to lead to different scores between the twins. Note that the prompt says the twins have the same "type of job" and "same salary". However they could have these characteristics while working in different sectors. Suppose instead of the same 'type' of job, the twins have the exact same job to remove ambiguity. Furthermore, let us consider both twins as statisticians (or data scientists, or any other term used to describe the role). Statisticians are ubiquitous – they are all over different industries, hence the twins could easily work in different sectors. Though the Irish data is not clear on what it means by 'Industry sector', let us take it to mean that a person working in any capacity for "Amazon" would be part of the retail sector, while working for a financial company counts as working in the financial sector. Using the Irish dataset sectors, one could work in retail (e.g. as a demand forecaster), while one could work in financial services (e.g. in market risk). Comparing the bad

rate for these attributes, we have 8% for retail services, and 5% for financial services. Hence despite working the same type of job with the same salary, the twins could be priced differently because of the sector they work in. The tables to obtain this information are in **Appendix Q4**.

For **gender**, although it is possible there is a gender split in risk, it is likely the twins are the same gender. To show this, assume monozygotic (identical) and dizygotic (fraternal) have probabilities of 1/3 and 2/3 respectively (McAslan Fraser, 2012). Identical twins are always the same gender. Within fraternal twins, consider each combination (MM, FF, MF, FM) equally likely. Then we have a 50% chance twins are the same gender given they are dizygotic. Hence the probability of twins being the same gender is:

$$P(Same) = P(Same|Mono)P(Mono) + P(Same|Dizy)P(Dizy)$$
$$P(Same) = 1\left(\frac{1}{3}\right) + 0.5\left(\frac{2}{3}\right) = \frac{2}{3}$$

Additionally, it may be illegalto ask for a customer's gender. We do not know the regulations the bank is working under, so it could be illegal to score on gender. Additionally, a bank could be trying to remove gender from its scoring in **anticipation** of a regulation change on gender (for example, in 2012 it was made illegal to gender-price in insurance in the UK), so avoids including gender in a model.

Gender can of course be changed, however the proportion of the population who actually change their gender is very low, so it is unlikely this is a reason for the difference between the twins.

Question 5: Irish Data – Coarse Classing

Before beginning, we take note of the fact that the data has a clear order (1 transaction on average is greater than 0 transactions on average). So it is common sense not to group around numbers, for example grouping 0, 24-50, and 70-90 transactions does not make sense as there is a clear measurable distance between people in this dimension. However if considering marital status, if 'Single' and 'Divorced' behave similarly (i.e. similar bad rates), there is no quantitative assignment to either status (i.e. we cannot say 'Single' is "more" than 'Divorced'), so it is permitted to group them as similar.

The first step in coarse classing is to get the data in a suitable format. To do this, I copied the entire table in "Data 2010 Apps Only" into a new workbook, and created a pivot table. This pivot table had rows denoting the number of transactions in the last 6 months, with columns indicating whether a loan was categorized as good or bad. The elements in the pivot table were just the counts of each loan category. To illustrate this look at Table 2. The value of 18 shows that for people with exactly 2 transactions in the past 6 months, 18 of the loans made were bad (in this dataset). Note that this table is only the top few rows of the pivot table.

Table 2: Sample of Pivot Table

Count of CA Average Number of Transactions L6M			
Number of Transactions	**B**	**G**	**Grand Total**
0	25	223	248
1	13	50	63
2	18	51	69
3	21	55	76
4	17	64	81
5	9	57	66

As suggested, bands will have increments of 5 to begin with. Since the data becomes more and more sparse as the number of transactions gets high, we will reach a point where the band will be "X or more". This is to avoid attributes with 0 good or bad (or both) observations. We also note that the data is zero-inflated. That is, there are many zero values in the data (many people have made no transactions in the last 6 months). In fact, 0 is the most frequent average number of transactions in the last 6 months. We could suggest that people who make 0 transactions are substantially different to those who average even 1 transaction. For example, it could suggest a largely cash-based lifestyle. Such people would not have much money in their current account, which we can check in the data by looking at "Average Current Account Balance (L6M)". Plotting the log average balance against the average number of transactions yields Figure 1. We use the log average balance since the average balances can get very high, preventing us from seeing discrepancies.

Figure 1: Balance vs Transactions

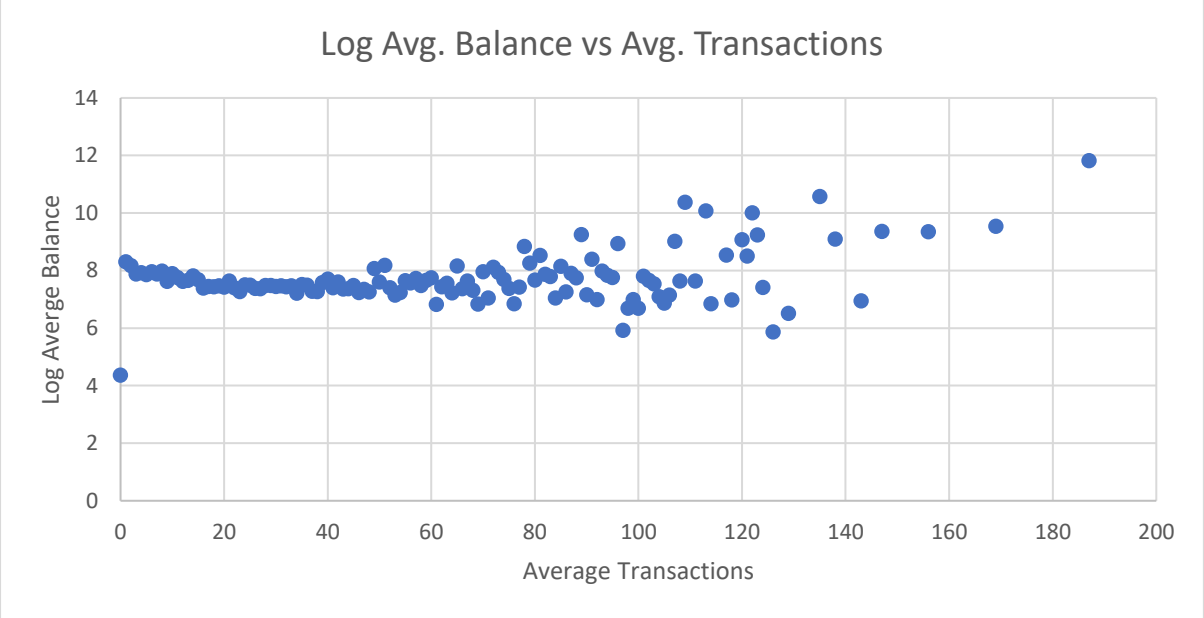
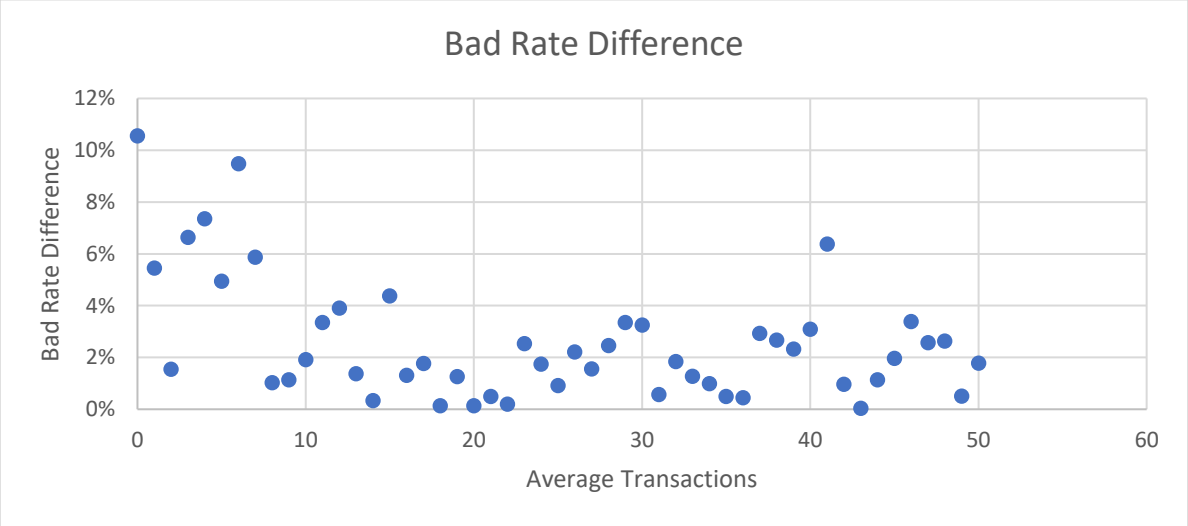


Figure 2: Bad Rate Difference vs Transactions



We see that for those with 0 average transactions, the log average balance is clearly the lowest, and is completely different to those who make even 1 transaction a month on average. If we retrieve the raw value, the average balance is in fact only £78. Additionally, Figure 2 shows the **difference** in the bad rate between a transaction average and its neighbour. For example, a value of 10% at 0 means the difference in bad rates between average transaction values of 0 and 1 is 10%. Figure 2 shows that the bad rate difference between 0 and 1 is the highest amongst average transaction values from 0 to 50 (going too much higher means sparse data, but the pattern is the same). Consequently, I decide to use '0' as its own band at the bottom of the coarse classing. The results can be seen in Table 3 along with the decided upon highest band. The construct of 'Lower' and 'Upper' columns is to allow the use of SUMIFS for 'Bads' and 'Goods' columns. This is carried forward when calculating the power, but the columns themselves are removed for aesthetics in future. To create the 'band' column in Tables 4 through 10 I concatenated the 'Lower' and 'Upper' columns with a hyphen between them.

Table 3: Determining Initial Classing

Lower	Upper	Bads	Goods	Count	Proportion
0	0	25	223	248	3.8%
1	6	84	340	424	7%
7	12	63	420	483	7%
13	18	69	699	768	12%
19	24	69	919	988	15%
25	30	55	846	901	14%
31	36	32	717	749	12%
37	42	36	558	594	9.2%
43	48	19	403	422	6.5%
49	54	18	266	284	4.4%
55	60	6	194	200	3.1%
61	66	4	111	115	1.8%
67	187	15	283	298	4.6%

Although the "61-66" band only has 1.8% of total loans (and only 4 bad loans), I decide to leave it in as a starting point, since it is easier to reduce the number of attributes than increase them. Similarly for the "55-60" band. This is because we can group attributes based on statistics such as weight of evidence, but doing the reverse is difficult since we cannot directly see where to cut an attribute. Hence it is advisable to have more bins than you might need as an initial step. The results of using the given bands from Table 3 are seen in Table 4.

Table 4: Results from Initial Coarse Classing

Band	**Bads**	**Goods**	**%Bads**	**%Goods**	**WoE**	**Information Value**	**ΔWoE**
0-0	25	223	5%	4%	-0.30	0.00	
1-6	84	340	17%	6%	-1.09	0.12	0.79
7-12	63	420	13%	7%	-0.59	0.03	0.50
13-18	69	699	14%	12%	-0.18	0.00	0.42
19-24	69	919	14%	15%	0.10	0.00	0.27
25-30	55	846	11%	14%	0.24	0.01	0.14
31-36	32	717	6%	12%	0.62	0.03	0.38
37-42	36	558	7%	9%	0.25	0.01	0.37
43-48	19	403	4%	7%	0.56	0.02	0.31
49-54	18	266	4%	4%	0.20	0.00	0.36
55-60	6	194	1%	3%	0.98	0.02	0.78
61-66	4	111	1%	2%	0.83	0.01	0.15
67-187	15	283	3%	5%	0.45	0.01	0.39

The %Bads and %Goods columns indicate what proportion of the total bads/goods are in a particular band. The weight of evidence is calculated using $WoE_i = LN\left(\frac{Goods_i}{Bads_i} * \frac{Bads}{Goods}\right)$, where i denotes the band. So the weight of evidence for band i is the natural log of the good to bad ratio in that band, multiplied by the bad to good ratio for the population. The penultimate column is summed to obtain the information value ($IV = \sum_i WoE_i * (\%Good - \%Bad)_i$). Note that the column name is just for clarification, and each individual value in the column is not itself an information value. The final column shows the difference in the weight of evidence between adjacent bands. This will be used later to reduce the number of bands by grouping bands with similar weights of evidence.

Using the penultimate column, we found the power (1000*IV) to be 267.55. Hence this characteristic is capable of being used in the modelling phase, if we use the rule of thumb that a power of greater than 200 is suitable.

Simplification

Note that this was a starting point, and now we aim to simplify the model by reducing the number of attributes. Doing this increases the bias of the model (we do not distinguish as well between different attributes) but reduces the variance (results carry less uncertainty as we have more data in each attribute). Essentially, we trade off overfitting for underfitting.

Two clear candidates for high variance in the results are the "55-60" band, and the "61-66" band due to their low level of bad customers. As a result, we combine these bands together, reinforced by the fact the weight of evidence disparity is only 0.15, the second lowest in the available attributes. The lowest weight of evidence disparity is for the "19-24" and "25-30" bands, hence we group these attributes together as well. Maintaining increments of 5 around the simplifications yields Table 5.

Table 5: Simplified Coarse Classing

Band	Bads	Goods	%Bads	%Goods	WoE	Information Value	ΔWoE
0-0	25	223	5%	4%	-0.30	0.00	
1-6	84	340	17%	6%	-1.09	0.12	0.79
7-12	63	420	13%	7%	-0.59	0.03	0.50
13-18	69	699	14%	12%	-0.18	0.00	0.42
19-30	124	1765	25%	30%	0.16	0.01	0.34
31-36	32	717	6%	12%	0.62	0.03	0.45
37-42	36	558	7%	9%	0.25	0.01	0.37
43-48	19	403	4%	7%	0.56	0.02	0.31
49-54	18	266	4%	4%	0.20	0.00	0.36
55-66	10	305	2%	5%	0.93	0.03	0.72
67-187	15	283	3.0%	4.7%	0.45	0.01	0.48

Doing the same calculation as previously, we find that the power has only reduced slightly, to 265.95. Consequently, it would be preferable to use this simplified coarse classing, as we reduce variance with largely no drop in power (i.e. drop in bias). However, the "55-66" band still contains only 10 bads. Hence, if we further wish to reduce the variance we can combine it with "67-187", although the weight of evidence difference suggests these attributes are very different. Additionally, we could combine "43-48" with "49-54" if we wish to obtain an approximate minimum of 20 bads for each attribute. Applying both these simplifications yields Table 6.

Table 6: Second degree simplified coarse classing

Band	**Bads**	**Goods**	**%Bads**	**%Goods**	**WoE**	**Information Value**	**ΔWoE**
0-0	25	223	5%	4%	-0.30	0.00	
1-6	84	340	17%	6%	-1.09	0.12	0.79
7-12	63	420	13%	7%	-0.59	0.03	0.50
13-18	69	699	14%	12%	-0.18	0.00	0.42
19-30	124	1765	25%	30%	0.16	0.01	0.34
31-36	32	717	6%	12%	0.62	0.03	0.45
37-42	36	558	7%	9%	0.25	0.01	0.37
43-54	37	669	7%	11%	0.40	0.01	0.15
55-187	25	588	5%	10%	0.67	0.03	0.26

Again calculating the power via 1000*IV returns a power of 258.71. We have again reduced the attributes without impacting the power too much, and we in fact notice that "37-42" and "43-54" have a similar (0.15 difference) weight of evidence, so we may be able to further simplify. We get Table 7 doing this.

Table 7: Third degree simplified coarse classing

Band	**Bads**	**Goods**	**%Bads**	**%Goods**	**WoE**	**Information Value**	**ΔWoE**
0-0	25	223	5%	4%	-0.30	0.00	
1-6	84	340	17%	6%	-1.09	0.12	0.79
7-12	63	420	13%	7%	-0.59	0.03	0.50
13-18	69	699	14%	12%	-0.18	0.00	0.42
19-30	124	1765	25%	30%	0.16	0.01	0.34
31-36	32	717	6%	12%	0.62	0.03	0.45
37-54	73	1227	15%	21%	0.33	0.02	0.29
55-187	25	588	5%	10%	0.67	0.03	0.34

Calculating the power now returns 257.66, so we have simplified again with only a small loss of power. Overall, with only 3 simplifications we have reduced the power by only 10 points whilst reducing the number of attributes from 13 to 8. For me, if it was a concrete choice between the above classings, I would select Table 7, as there are a good number of observations in each attribute in comparison to the levels of higher power, which suggests the results are more reliable.

As a final step, we alter the increment (in the bands which still use increments, e.g. "1-6") to try and obtain a higher level of power. Altering the increment to 4 yields Table 8.

Table 8: Third degree simplification with increment 4

Band	Bads	Goods	%Bads	%Goods	WoE	Information Value	ΔWoE
0-0	25	223	5%	4%	-0.30	0.00	
1-5	78	277	16%	5%	-1.22	0.14	0.92
6-10	46	311	9%	5%	-0.58	0.02	0.64
11-15	62	494	13%	8%	-0.42	0.02	0.16
16-30	154	2142	31%	36%	0.14	0.01	0.56
31-35	26	609	5%	10%	0.66	0.03	0.52
36-54	79	1335	16%	22%	0.34	0.02	0.33
55-187	25	588	5%	10%	0.67	0.03	0.33

The power is now 274.25, whilst maintaining a reasonable amount of loans in each band. This is in fact an improvement on the initial power of 267.55. As a result, thanks to the initial process of simplification while maintaining an increment of 4, we have been able to simplify **and** obtain a higher level of power than the case with 13 attributes.

At this point, it would still be acceptable to use the data for modelling, as each attribute has a good number of loans. It may also be desirable to simplify if we wish to further reduce the variance, and the "6-15" band is a good candidate due to the minimal change in weight of evidence. Additionally, 25 bads may not be a convincing number for higher management, so we may use "31-187" as a band. However I would advise against combining the 0 band with anything, due to the wildly different behaviour compared to low transactions (as can be seen in the weight of evidence difference). For completeness, I will report the tables for the above scenarios. Table 9 shows the "31-187" grouping (with increment 4), and Table 10 shows the impact of altering the lowest band to include 1.

Table 9: "31-187" grouping with increment 4

Band	Bads	Goods	%Bads	%Goods	WoE	Information Value	ΔWoE
0-0	25	223	5%	4%	-0.30	0.00	
1-5	78	277	16%	5%	-1.22	0.14	0.92
6-10	46	311	9%	5%	-0.58	0.02	0.64
11-15	62	494	13%	8%	-0.42	0.02	0.16
16-30	154	2142	31%	36%	0.14	0.01	0.56
31-187	130	2532	26%	42%	0.48	0.08	0.34

The power in this scenario is 265.17. This is a 9 point loss, which seems quite large, however there is likely a large variance reduction by now having a large number of observations in each band. Only having 6 attributes vastly simplifies the model, and we still have a good degree of power.

Table 10: Table 8 with a '0-1' lower band

Band	Bads	Goods	%Bads	%Goods	WoE	Information Value	ΔWoE
0-1	38	273	8%	5%	-0.52	0.02	
2-6	71	290	14%	5%	-1.08	0.10	0.56
7-11	51	344	10%	6%	-0.58	0.03	0.50
12-16	59	503	12%	8%	-0.35	0.01	0.23
17-30	146	2037	29%	34%	0.14	0.01	0.49
31-35	26	609	5%	10%	0.66	0.03	0.52
36-54	79	1335	16%	22%	0.34	0.02	0.33
55-187	25	588	5%	10%	0.67	0.03	0.33

The power in this scenario is 250.35. This is still high, but just the slight alteration of including 1 in the lower band has reduced power by 14 points. However if higher management were insistent on having a larger band, the power is still high enough to be used for modelling.

Question 6: "Previous Loan" Characteristic

We are judging this characteristic based on whether it is suitable to include in a scorecard ranging from 100 to 700.

In terms of the characteristic as a whole, we would expect (ceteris paribus) previous loan behavior to continue with a new loan. That is, assuming there have been no major changes in someone's life, there is no reason to suspect their attitude towards loan repayment would change. Hence it makes intuitive sense to use this characteristic.

Although there is not data available to test, intuitively we would expect the characteristic to have good differentiating properties, for the reasons above.

With regards to legality, the process seems fine as the lender largely uses their own files. Additionally, data from the credit reference agency is gathered via public record and by people consenting to a credit check.

The characteristic is also stable, as their previous loan status will not change after being provided the loan. Although their new performance may be different, that is more due to the inherent risk of providing a loan.

The data on the characteristic itself seems simple enough to obtain, however the reality of messy real world data perhaps means that the applicant's details being "matched against the lender's files" is a complicated data cleaning job. Though without further information, this is pure speculation based on language choice.

The primary weakness of this characteristic is the scores assigned when compared to the range of the scores. Assuming all characteristics had similar scores, we would require around 35 characteristics at the highest score to achieve a maximum score (600/17). Of course this is assuming the best on each characteristic, however if we are setting 700 as a *theoretically* feasible score, then we should be able to achieve it with our given attributes. The point of this analysis is to say that the characteristic seems largely insignificant in the grand scheme of the score, despite intuitively being a very important characteristic in that it describes the exact behaviour we want to predict. As such, despite being *intuitively* differentiating between good and bad, the low scores suggest this characteristic does not actually differentiate well between good and bad loans.

Additionally, the (ceteris paribus) difference in scores between someone with experience of repaying a loan with no missed payments (17) and someone who did not repay their loan (9) is very small relative to the range of the scores. We would perhaps expect there to be a greater difference, since we are comparing a good chance of a fully repaid, good quality loan to a good chance of not being repaid.

Lastly, there is a problem in the attribute with regards to 'some' missed payments. Within this attribute, a single missed payment is an equivalent level of risk as someone who had a habit of missing payments. It is clearly worrying that someone suffering a temporary negative shock to

their finances is classed the same as someone who has a habit of missing payments. These are clearly 2 very differently behaving people, and we know intuitively that in the long-run the latter customer will be more likely to fail than the former. Hence if possible, it would be advisable to coarse class the characteristic with the number of missed payments. Of course this is dependent on the amount of data available, but if the data is indeed sufficient then this attribute choice is poor.

Bibliography

McAslan Fraser E. 2012. *Expecting twins, triplets or more: the healthy multiple pregnancy guide*. Tamba (Twins & Multiple Births Assocation). www.tamba.org.uk [Accessed February 2019].

Appendix

All data comes from the Irish dataset, in "Data 2010 Apps Only". To avoid information overload, I will include only the top few rows for large tables to provide a feel for the data. To denote these sample tables, I will write "(sample)" in the section names.

Q4

Information for Retail vs Financial Services

The first 4 columns come from a pivot table with rows "Industry Sector", columns "Status – G/B/R", values "Count of Status – G/B/R" with rejections filtered out. The sector name was found using the lookup table found on the next page.

Sector Code	**B**	**G**	**Grand Total**	Sector name	Bad Rate
A	14	134	148	Agriculture	9%
C	56	293	349	Construction	16%
DF	8	128	136	Defence Forces	6%
E	5	107	112	Energy and Utilities	4%
ED	17	293	310	Education	5%
EL	10	111	121	Entertainment and Leisure	8%
ET		2	2	Extra-territorial organisations and bodies	0%
FI		2	2	Fishing	0%
FS	23	440	463	Financial Services	5%
HR	15	115	130	Hotels and Restaurants	12%
HS	25	510	535	Health and Social Work	5%
M	35	760	795	Manufacturing	4%
MQ		12	12	Mining and Quarrying	0%
O	135	1067	1202	Other Services	11%
PH	1	1	2	Private Households with employed Persons	50%
PP	13	151	164	Professional Practice	8%
PS	36	656	692	Public Services	5%
RS	63	768	831	Retail/Wholesale Services	8%
T	21	305	326	Transport	6%
X	18	124	142	Unknown / Missing	13%

Lookup Table

Industry Sector	Description
A	Agriculture
C	Construction
DF	Defence Forces
E	Energy and Utilities
ED	Education
EL	Entertainment and Leisure
ET	Extra-territorial organisations and bodies
FI	Fishing
FS	Financial Services
HR	Hotels and Restaurants
HS	Health and Social Work
M	Manufacturing
MQ	Mining and Quarrying
O	Other Services
PH	Private Households with employed Persons
PP	Professional Practice
PS	Public Services
RS	Retail/Wholesale Services
T	Transport
X	Unknown / Missing

Q5

Data for Tables 2 – 10 (Sample)

The fields are “CA Average Number of Transactions L6M” (rows), “Status – G/B/R” (columns). “Count of CA Average Number of Transactions L6M” (values), and the rejected applications were filtered out. This was also used to see that 0 was the mode of the data.

Count of CA Average Number of Transactions L6M	**Status**		
L6M Transaction	**B**	**G**	**Grand Total**
0	25	223	248
1	13	50	63
2	18	51	69
3	21	55	76
4	17	64	81
5	9	57	66
6	6	63	69
7	12	54	66
8	8	57	65
9	10	65	75
10	10	72	82
11	11	96	107
12	12	76	88
13	11	102	113
14	13	104	117
15	15	116	131
16	8	105	113
17	12	131	143
18	10	141	151
19	10	138	148
20	15	172	187

Data for Figure 1 (Sample)

This was created using a pivot table. The rows were “CA Average Number of Transactions” while the values were “Average of CA Average Balance L6M”. This formed the first 2 columns. The final column used the “LN” formula on the “Average of CA Average Balance L6M” column.

Average Monthly Transactions (L6M)	Average of CA Average Balance L6M	Log balance
0	£79	4.37
1	£4,025	8.30
2	£3,536	8.17
3	£2,631	7.88
4	£2,741	7.92
5	£2,578	7.85
6	£2,849	7.95
7	£2,640	7.88
8	£2,918	7.98
9	£2,043	7.62
10	£2,677	7.89
11	£2,323	7.75
12	£2,043	7.62
13	£2,118	7.66
14	£2,479	7.82
15	£2,163	7.68
16	£1,612	7.38
17	£1,709	7.44
18	£1,692	7.43
19	£1,742	7.46
20	£1,672	7.42

Data for Figure 2 (Sample)

The first 4 columns were constructed by a pivot table. The fields were "CA Average Number of Transactions L6M" (rows), "Status – G/B/R" (columns), and "Count of CA Average Number of Transactions L6M" (values). A filter was applied so only accepted loans were in the data. The bad rate calculation is simply the 'B' column divided by the grand total. The bad rate difference of the ith row is $ABS(BadRate_i - BadRate_{i+1})$.

L6M Transactions	B	G	Grand Total	Bad Rate	Bad Rate Difference
0	25	223	248	10%	11%
1	13	50	63	21%	5%
2	18	51	69	26%	2%
3	21	55	76	28%	7%
4	17	64	81	21%	7%
5	9	57	66	14%	5%
6	6	63	69	9%	9%
7	12	54	66	18%	6%
8	8	57	65	12%	1%
9	10	65	75	13%	1%
10	10	72	82	12%	2%
11	11	96	107	10%	3%
12	12	76	88	14%	4%
13	11	102	113	10%	1%
14	13	104	117	11%	0%
15	15	116	131	11%	4%
16	8	105	113	7%	1%
17	12	131	143	8%	2%
18	10	141	151	7%	0%
19	10	138	148	7%	1%
20	15	172	187	8%	0%