2023-24

CMSE11428: Predictive Analytics and Modelling of Data

Credit Score Showdown: Unveiling the Power of Predictive Models

GROUP 7

Wor d count: 3.890

Executive Summary
Banking is a massive industry, and one which requires a great deal of risk management. This is particularly prevalent in credit, which is why the need to accurately measure customers credit risk is a key problem for all banks. This paper will highlight some methodologies that can be used in estimating this risk based on the dataset provided. Firstly, we analyse a random forest model. This is a strong predictive model with a high potential in credit risk as it creates a highly accurate and robust model for calculating credit. Our random forest model returns an accuracy of 86% for measuring whether a new customer is correctly classified. Although we would recommend the use of this model, regulations will likely not permit it due to its lack of transparency.

Therefore, the paper will also look to utilise a logistic regression model as it is easy to interpret, however we will enhance our logistic regression with the use of the Weight of Evidence (WOE) measure; a common addition used in the banking sector. The results from our logistic regressions with WOE gave an accuracy measure of 73%. However, this suffered from imbalance and thus a second regression with imbalanced handling resulted in a 56% accuracy score. We put a great deal of emphasis on this model as it is easily interpretable and as such complies with Brazil's financial regulations that mandate non-discrimination and demonstrable proof of it.

Section 1: Problem Statement & Introduction
On a day-to-day basis it is common practice for a bank to deal with risk and uncertainty. In fact, a bank's ability to deal with risk successfully directly impacts its performance and profitability (Torosyan, 2017). Credit risk is one of these common practices that banks across the world seek to deal with. There is a need to accurately represent a customer's credit risk so that the institution can make the most informed decision possible. This ensures that a bank does not give credit to a risky person, who is likely to default and thus, lose the bank money. This is done through the practice of credit scoring. Credit scoring models analyse information given by the customer including income, assets, previous credit history etc, to assess the likelihood that this individual may default on their credit; measured through a score which indicates whether they are indeed risky. If a bank cannot be accurate in measuring risk, the bank will actively misinterpret customers looking for credit, and consequently will face losses. Considering the competitive nature of the financial and banking industry across the world, it is necessary to gain advantages in credit scoring to ensure you as a bank do not fall behind. This paper will aim to give you recommendations on how best to improve the accuracy of your credit scoring models to help mitigate uncertainty, and lower the risk taken when providing credit to customers and thus, improve profitability.

Section 2: Literature Review & Model Justification
Using statistics to measure 'credit scores' quickly became common by as early as the 1970's (Dumitrescu *et al*., 2020). It was from then that the idea of logistic regression came to the forefront. While there are many other ways of measuring risk, logistic regression is still commonly used even today in credit scoring. The reasons for this include how easily interpretable the results are and the fact that there is no need for parameter tuning (Szepannek, 2017).

However, there is clear problem with standard logistic regression in that it requires linearity between variables. In practice, this is unlikely with non-linear relationships common between variables. Higher dimensional data makes this also more likely to occur. For example, it has been shown that there is a non-linear relationship between age and default rate (Szepannek, 2017). Szepannek then states that the logistic regression would not be appropriate. Moreover, arguments for ensemble methods are based on this idea that they outperform logistic regression due to the non-linearity of the data (Dumitrescu *et al*., 2022). Many authors show outperforming machine learning models compared to logistic regression including support vector machines (Van Gestel *et al*., 2003), decision trees (Westreich *et al*., 2010), neural networks (Bensic *et al*., 2005) and random forest (Do and Simioni, 2021).

Lessman et al. (2015) recommend that the benchmark classification technique should be random forests for comparison to future advances rather than the standard logistic regression as after a comprehensive

review of the literature they conclude that ensemble methods in general outperform logistic regression. The random forest approach has gathered favour due to the fact that it is random in nature (Ghatasheh, 2014) which allows for reduced variance as a result of a decorrelation between trees. Furthermore, the added benefits of an algorithm like random forest is that it massively decreases the time and money spent on preprocessing the data as it is very robust in handling missing values (Risk Advisory, 2019). Overall, it is common for the literature to use random forests as they are one of the most accurate tools when coming to assessing default risk and consumer credit. Therefore, out of the ensemble methods we choose to provide analysis of the random forest model, as it clearly is one of the best performing models in the credit risk literature.

There have also been advances in the logistic regression. For example, Dumitrescu et al. (2022) improve logistic regression with non-linear decision-tree effects where the results become competitive with random forest. Moreover, commonly used in the banking sector is the Weight of Evidence (WOE), which can be applied to logistic regression. Chen., *et al* (2020) use the WOE measure to accurately calculate potential defaulters, stating that WOE is incredibly beneficial in the face of missing or very high dimensional data. An application of over 100,000 datapoints with many missing values show that the logistic regression with WOE outperformed a typical logistic regression model. Similar findings were also found by Yuan (2018) who assessed credit scoring in the peer-to-peer loan system in China. Their results conclude also that the introduction of the WOE improves the accuracy of their logistic regression model. Therefore, in our model we seek to implement this WOE measure to also improve accuracy particularly with many missing values.

While it is clear from the literature that in general machine learning ensemble methods tend to outperform logistic regression in ROC and accuracy, these methods encounter the 'black box' problem. This is where we cannot interpret the importance of variables and only see the output; causing a lack of transparency (Bücker *et* al., 2022). In finance and banking this becomes problematic with consumer safety a key regulatory measure. In fact, in 2018 Brazil updated their data protection law in which article 6 part IX includes an update to non-discrimination of consumers (Lemos *et al*., 2018). Mendes and Mattiuzzo (2022) note that this change may directly impact credit scoring, although it is yet to be seen in practice how it will affect the industry. With this in mind, logistic regression may be the most viable option due to the regulatory nature of credit from the main public-body and the lack of interpretability (and potential discrimination) that black box methods may incur.

Due to our review of the literature, we provide a comparison of two types of models. Logistic regression with WOE and random forest. However, we also compare the logistic regression with WOE, with and without imbalanced handling. All models are compared to a standard logistic regression as a baseline. We use both as we do not know exactly how the updated consumer protection law is acted upon in Brazil, and thus, expect that ensemble methods may not meet regulatory expectations.

Section 3: Data Pre-Processing

3.1 Variables

The dataset comprises of 54 features and 50,000 datapoints. Each variable underwent a detailed analysis for inclusion in our models. Variables creating unnecessary noise were excluded. 'SEX' was omitted to non-discriminatory financial regulations. Discriminatory variables like 'AGE' were binned to comply with regulatory standards. A standard correlation matrix assessed variable correlation, ensuring adherence to the no multicollinearity assumption in logistic regression.

3.2 Missing Values

There were several variables that containing many missing values. 'PROFESSIONAL PHONE AREA CODE' had more than 70% missing values and this variable was dropped as such. If included there would be inherent bias in the results as replacement of the data may not be representative of the

population at all. It is important to note WOE deals with missing values and such this process is purely precautionary.

Furthermore, other variables indicated problematic data collection. 'QUANT ADDITIONAL CARDS' displayed zero variance, suggesting that this data was collected incorrectly. Other anomalies noticed were negative values included in variables such as 'PAYMENT_DAY.' These we believe had incorrect information and thus were not included in our analysis.

3.3Outliers

We also run an outlier analysis on our remaining variables of choice. Rousseeuw & Hubert (2011) note that these severe datapoints can impact statistical inference and bias the results. We use the IQR box plot method to decipher whether any of the remaining variables do indeed include outliers. Many variables due to their categorical nature did not show outliers. However, 'AGE' did include outliers. It was noted that according to the data a six-year-old, a seven-year-old, 14 year-olds and 16 year-olds all applied for credit and were considered outliers. When considering options, we looked to more data from Brazil and note that in the past 16-year-olds have successfully applied for credit. Thus, we removed those under 16 from our dataset as we believe this unreasonable.

3.4 Imbalance

Next, we turn our attention to the target variable. We note that the target variable is imbalanced with 76% of customers regarded as non-defaulters, and 24% regarded as defaulters. In credit scoring particularly, if class imbalance is not considered, this can lead to a bias of results towards the good creditors (Brown & Mues, 2012) (as this class has the higher value of customers in the dataset). Therefore, the accuracy measures of the defaulters will result in being substantially lower than expected. As a result, we use Random Over Sampling which generates more examples of the minority target category and helps balance the dataset. This will lead to a precise and consistent model in handling both classes. We provide a comparison logistic regression with and without imbalance handling.

3.5 The Weight of Evidence (WOE)

After completing the data preprocessing, we applied the Weight of Evidence as a binning method, as numerous studies have emphasized its benefits in credit scoring (Persson, 2021).

WOE categorizes variable values, assigning values based on the proportion of defaulters to non-defaulters. When a category has a higher proportion of defaulters, the WOE value becomes significant, indicating effective separation (Persson, 2021). Utilizing WOE requires binning.

Binning involves grouping either continuous or categorical variables into distinct sets or bins. As outlined by Persson, (2021), an effective binning strategy should adhere to the following principles:

- Treat missing values as a separate group.
- Ensure that each bin comprises at least five percent of all observations.
- Avoid having bins with zero instances of either good or bad loans.

The calculation procedure is conducted as follows. Consider a dataset comprising N independent observations, where Y represents a binary dependent variable with values 1 for default and 0 for not default. Let X1, X2, ..., Xp denote a set of independent variables, and B1, B2, ..., Bk represent bins for the variable Xj. The Weight of Evidence (WOE) for the variable Xj in bin i is subsequently defined as: (Persson, 2021)

$$Y = \{\frac{1\ if\ default}{0\ if\ the\ not\ default}\}$$

$$WOE_{ij} = log\frac{P(X_j \epsilon B_i \,|Y = 1)}{P(X_j \epsilon B_i \,|Y = 0)}$$

The Weight of Evidence (WOE) quantifies the effectiveness of grouped attributes in differentiating defaulters from non-defaulters. A high negative WOE value corresponds to a greater risk of default, whereas a high positive value indicates a lower risk of default (Persson, 2021).

The Weight of Evidence (WOE) approach comes with strong advantages. WOE is designed to establish a monotonic relationship with the dependent variable. It is noted that while non-monotonic relationships can occur, they can still be acceptable as long as the relationship can be adequately explained. Additionally, WOE is adept at handling missing values and outliers, providing a convenient solution in such situations (RPubs, Rstudio, March 2018). Hence, we believe WOE can improve the accuracy of out model.

3.6 Random Forest – Extra Pre-processing step

The random forest model underwent an extra pre-processing step. After determining the characteristics of the features, a Column Transformer was utilized to implement a series of transformations tailored to the data types, alongside one-hot encoding categorical variables into dummy variables. The strategy adopted to contend with missing values in numerical columns was the imputation of the mean, a technique that replaces missing data with the average value, thus preserving the overall distribution.

Section 4: Analysis & Results

4.1 Random Forest

Firstly, We present the robust results of our Random Forest model in table 1. During model selection, Random Forest highlighted influential features, including 'PERSONAL_MONTHLY_INCOME' and 'PROFESSION_CODE,' indicating their significance as predictors aligned with expectations of financial stability and occupational types influencing credit risk

Table 1: Random Forest Results

	Precision	**Recall**	**F1-score**	**Support**
Good Class (0)	.87	.84	.85	11,064
Bad class (1)	.84	.87	.86	11,111
Accuracy			.86	22,175

The application yields precision rates of 87% for class 0 and 84% for class 1, demonstrating high accuracy in credit loan predictions. The recall is also robust for both classes, indicating the model's consistency in achieving similar results. The well-balanced data contributes to a strong overall accuracy rate, ensuring reliable predictions for both classes.

In summary, the Random Forest model excels in accurately predicting credit risk, effectively distinguishing between defaulting and non-defaulting customers. Despite its high predictive power, its lack of interpretability poses challenges in regulatory compliance.

4.2 Logistic Regression with WOE

For the logistic regression, the analysis utilizes pre-processed data, with the variable transformations done using the Weight of Evidence (WOE) approach. We then provide a comparison to a standard logistic regression on performance.

The pre-processing involved categorizing variables into groups using the WOE method. This entailed classifying variables based on the similarity between observations for each value and its weight of evidence. Subsequently, a logistic regression model was fitted to the transformed data. With the final variables shown in Table 2. (see appendix A1 for the detailed list).

Table 2: Final Variables

COMPANY	PRODUCT
MARITAL STATUS	POSTAL ADDRESS TYPE
MONTHS IN RESIDENCE	RESIDENCE TYPE
MONTHS IN THE JOB	RESIDENCIAL STATE
OCCUPATION TYPE	EDUCATION LEVEL
PAYMENT DAY	AGE
PERSONAL MONTHLY INCOME	

The logistic regression model is expressed by the following equation:

$$Probability\ of\ default = \frac{1}{1 + e^{(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}}$$

where:

- Probability of Default: Represents the likelihood of a credit default occurring.
- β0, β1 and βn: Coefficients associated with each predictor variable
- X1…. Xn : Predictor variables derived from the pre-processed data.
- *e*: The base of the natural logarithm.

The logistic function (sigmoid) constrains the output to the range [0, 1], mapping the linear combination of predictors to a probability:

$$Log - odds = 1 + e^{(\beta_0 + \beta_{0+} + \cdots + \beta_n \beta_n)}$$

4.3 Variable Transformation using WOE

In the context of the Weight of Evidence (WOE) transformation:

- $WOE = \ln(\frac{\%\ Non-Events}{\%\ Events})$
- In this model class (0) means default and class (1) non-default.
- % Non-Events: Percentage of non-default instances in a particular category.
- % Events: Percentage of default instances in the same category.

The WOE is used to transform categorical variables into continuous variables, enabling logistic regression to be applied. It helps capture the relationship between the predictor variables and the log-odds of credit default.

4.4 Reference Categories:

Reference categories are crucial in logistic regression, serving as the baseline for comparison. The choice of reference categories influences the interpretation of coefficients. The reference categories are shown in table 3. The reference categories were chosen through the WOE transformation as the WOE calculated which category had the most likely defaulters.

Table 3: Reference Categories

REF_CATEGORIES:	
COMPANY: Y	PRODUCT: 7
MARITAL: 1	POSTAL_CODE: 2
MONTHS_IN_RESIDENCE: >60	RESI_TYPE_ 2
MONTHS_IN_THE_JOB: <0	STATE_RESI: SE
OCC_TYPE: 4.0	EDUCATIONAL :0.0
PAYMENT_DAY: 25	AGE: >83
PERSONAL_MONTHLY_INCOME: 845-1238.872	

4.5 Model Coefficients and P-values

The logistic regression model calculates coefficients (β0, β1,,*βn*) for each predictor variable. The p-values associated with these coefficients indicate the significance of each variable in predicting credit risk. Lower p-values suggest a more significant impact on the model. (See appendix A2)

4.6 Analysis coefficients and p-values of age:

Table 4: Coefficient and P-values

Feature Names	**Coefficients**	**P_values**
Intercept	-0.736563	Nan
AGE: <16.9	-0.589677	0.67925747
AGE: 16.9-30	0.194936	0.10003358
AGE: 31-43.3	0.437502	0.00008785
AGE: 43-56.5	0.678415	0.00000000
AGE: 56-66.4	0.940607	0.00000000
AGE: 66-82.9	0.929220	0.00000000

From table 4 we show the relevant coefficient and p values for the age categories with the reference being 83+. The results show that for example being between 56-66.4, (since =1 is non-defaulter), is statistically significant in being associated with a higher log odd of being a good creditor when compared to being 83+. Therefore, we can interpret this as if one is 56-66.4 years old they are much more likely to be a good creditor than if they were 83+

- The age groups 31-43.3, 43-56.5, 56-66.4, and 66-82.9 seem to have a statistically significant positive impact on the log-odds of the outcome compared to the reference group.
- The age group <16.9 and 16.9-30 do not seem to have statistically significant impacts on the outcome at conventional significance levels.

Since one or more age subcategories exhibit statistical significance, we retain all these subcategories for inclusion in our model. Other strongly salient variables include Product, State of Residence and payment day. In future research we would magnify our look into these variables. Table A2 in the appendix shows the list of all coefficients and p-values.

4.7 Logistic Regression with WOE

It is important to note that for our logistic regressions we have swapped the target variable round so that Y=1 is a non-defaulter and Y=0 is a defaulter.

Table 5: Results for LR + WOE

	Precision	Recall	F1-score	Support
Bad Class (0)	.67	0	0	2,664
Good class (1)	.73	1.0	.85	7,336
Accuracy			.73	10,000

Table 5 shows the results for this model. The model performs well in predicting the positive class (Class 1) with high precision (0.73), recall (1.0), and F1-score. However, it performs poorly in predicting the negative class (Class 0) with low precision, recall, and F1-score. The accuracy is influenced by the imbalanced class distribution, and the weighted average F1-score provides a more balanced assessment of the model's overall performance.

4.8 Logistic Regression with WOE & Imbalance handling

Table 6: Results for LR + WOE + Imbalance Handling

	Precision	Recall	F1-score	Support
Bad class (0)	.56	.55	.55	7,417
Good Class (1)	.55	.57	.56	7,367
Accuracy			.56	14,784

The results from this model are shown in table 6. The imbalanced dataset has reduced the overall accuracy of the model but has made it slightly equal in precision. The precision rate for both defaulters and non-defaulters stands at 55% and 56% with a recall of 55% for defaulters and 57% for non-defaulters, (as we made = 1 the non-defaulters). While overall accuracy is lower, this may be more beneficial to you as a bank as it clearly is beneficial in both predicting defaulters and non-defaulters unlike the WOE without imbalanced handling which performs very poorly at capturing defaulters.

4.9 Comparative Analysis Results

Table 7: Model Comparison

Models	Accuracy Rate	AUC	GINI
LogR: Imbalance handling	0.50	0.49	-0.001
LogR: WOE	0.73	0.60	0.19
LogR: Imbalance handling, WOE	0.56	0.56	0.11
Random Forest	0.86	0.86	0.71

Table 7 represents the comparative performance stats of the 4 models. The first is a standard logistic regression with an imbalancing measure. The second and third are logistic regressions with WOE, the third including imbalancing checks. The fourth presents the results of our random forest model.

The random forest boasts an accuracy rate of 86% which is far superior to any other model. Yet, we believe that this model will not meet regulatory standards. A comparison of the logistic regressions shows that the WOE without imbalancing is the strongest predictor in terms of accuracy at 73%, however the dataset exhibits a significant class imbalance, particularly against the "Bad Class," leading to a recall of 0% for class(0) in the initial model. This imbalance may cause the model to prioritize learning patterns associated with the majority class, impacting its ability to identify instances of the minority class. In contrast, the imbalanced WOE model shows improvement with a recall of 55% for class(0) and 57% for class(1), addressing the initial model's limitations. Nevertheless, both models with WOE outperform just a standard logistic regression dramatically with an accuracy as good as just guessing at 50%.

Another common measure of model performance is the AUC score, which measure the area under the ROC curve. This is measured across all threshold levels. The AUC is designed to analyse the ability of the model to discriminate between bad and good outcomes (for example false positives and negatives).

Therefore, a greater AUC value indicates lower probabilities of type I and II errors. The AUC of the random forest model substantially outperforming other models with a score of 0.86, being able to consistently match the correct outcomes. The logistic regression with WOE for imbalanced and non-imbalanced datasets has similar scores with the non-imbalanced scoring 0.6 and with imbalancing scoring 0.56. Again, the standard logistic regression is as good as guessing with a score of 0.49. The gini index also measures performance and is derived from the AUC thus it follows similar results.

Section 5: Conclusion & Recommendations

Based on the results from our analysis we recommend to use the random forest model if regulation permits. It has very strong accuracy at 86% and a good precision at predicting outcomes of both defaulters and non-defaulters in terms of precision and recall. However, we understand the risk of relying on this model considering the impact of regulation. Hence, our second recommendation would be to use the logistic regression with WOE if random forest fails due to regulation. While not as powerful as the random forest model, it still outperforms a standard logistic regression while being completely interpretable thus, clearly ready for use. The decision to use imbalanced handling is personal however, for equality of prediction using WOE with imbalance will equally predict defaulters and non-defaulters. This extra performance gain can still be hugely important in such a competitive environment. We have provided a brief interpretation of one of the coefficients of the age variable. In future we recommend going into more detail of these coefficients to truly understand where you should focus your models. A list of strongly salient variables we believe should be focused on first include age, state_of_residence, product and payment_day.

To fortify credit scoring methodologies and especially using Logistic regression, it is recommended to include more financial measures including incorporating credit records, enriching the database with pertinent financial details, and integrating external scores. The inclusion of comprehensive credit records, encompassing payment history and credit utilization, provides a more comprehensive assessment of individuals' creditworthiness. Strengthening the database with additional information, such as financial behaviours and demographic details, enhances the model's analytical depth.

In conclusion, this paper has studied the literature around credit scoring in Brazil and has analysed two styles of models, random forest and logistic regression. We provide scores of the performance of these models while expressing their utility in the workplace. We conclude that random forest is the strongest predictor of credit risk, yet logistic regression with WOE seems more applicable to the work environment. In future to boost the performance of these models we would investigate a greater collection of data, alongside a focus on improving prediction of the logistic regression.

Bibliography:

Bensic, M., Sarlija, N. and Zekic-Susac, M., 2005. Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *13*(3), pp.133-150.

Brown, I. & Mues, C., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), pp.3446–3453.

Bücker, M., Szepannek, G., Gosiewska, A. and Biecek, P., 2022. Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, *73*(1), pp.70-90.

Chen, K., Zhu, K., Meng, Y., Yadav, A. and Khan, A., 2020. Mixed credit scoring model of logistic regression and evidence weight in the background of big data. In *Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 1* (pp. 435-443). Springer International Publishing.

Do, H.N. and Simioni, M., 2021. *A comparison of random forest and logistic regression model in credit scoring of rural households* (No. hal-03322462).

Dumitrescu, E., Hué, S., Hurlin, C. and Tokpavi S., 2020. Machine Learning or Econometrics for Credit Scoring: Let's Get the Best of Both Worlds. *SSRN Electronic Journal*. doi:https://doi.org/10.2139/ssrn.3553781

Dumitrescu, E., Hué, S., Hurlin, C. and Tokpavi, S., 2022. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. *European Journal of Operational Research*, *297*(3), pp.1178-1192.

Ghatasheh, N., 2014. Business analytics using random forest trees for credit risk prediction: a comparison study. *International Journal of Advanced Science and Technology*, *72*(2014), pp.19-30.

Lemos, R., Douek, D., Franco, S., Ramon, A.D., Santos, N. and Langenegger (2018). *Translation by*. [online] Available at: https://www.pnm.adv.br/wp-content/uploads/2018/08/Brazilian-General-Data-Protection-Law.pdf [Accessed 26 Nov. 2023].

Lessmann, S., Baesens, B., Seow, H.V. and Thomas, L.C., 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), pp.124-136.

Markov, A., Seleznyova, Z., & Lapshin, V. (2022). Credit Scoring Methods: Latest Trends and Points to Consider. *The Journal of Finance and Data Science*, 8, 180-201. https://doi.org/10.1016/j.jfds.2022.07.002https://doi.org/10.1016/j.jfds.2022.07.002

Mendes, L.S. and Mattiuzzo, M., 2022. Algorithms and discrimination: the case of credit scoring in Brazil. In *Personality and data protection rights on the internet: Brazilian and German approaches* (pp. 407-443). Cham: Springer International Publishing.

Persson, R. 2021, "Weight of evidence transformation in credit scoring models: How does it affect the discriminatory power?", Lund University & Ola Jönsson, Nordea.

RPubs, Rstudio, March 2018, "WOE, IV and Scorecards in Credit Risk Modelling", Available at: https://rpubs.com/erblast/creditrisk/

Risk Advisory. (2019). Available at: https://www2.deloitte.com/content/dam/Deloitte/sg/Documents/financial-services/sg-fsi-machine-learning-credit-risk.pdf.

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 73–79. https://doi.org/10.1002/widm.2

Szepannek, G., 2017. On the practical relevance of modern machine learning algorithms for credit scoring applications. *WIAS Report Series*, *29*, pp.88-96.

Torosyan, N., 2017. Application of binary logistic regression in credit scoring. *University of Tartu.*.' [online] Available at: https://core.ac.uk/download/pdf/85144517.pdf [Accessed 26 Nov. 2023].

Van Gestel, I.T., Baesens, B., Garcia, I.J. and Van Dijcke, P., 2003, January. A support vector machine approach to credit scoring. In *FORUM FINANCIER-REVUE BANCAIRE ET FINANCIAIRE BANK EN FINANCIEWEZEN*- (pp. 73-82). UNKNOWN.

Westreich, D., Lessler, J. and Funk, M.J., 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of clinical epidemiology*, *63*(8), pp.826-833.

Yuan, Z., 2018. Research on credit risk assessment of P2P network platform: Based on the logistic regression model of evidence weight. *Journal of Research in Business, Economics and Management*, *10*(2), pp.1874-1881.

Threshold: https://www.mathworks.com/help/risk/reject-inference-for-credit-scorecards.html

Appendix:

Appendix:

Table A.1 Imbalanced Variables

Features Name
marital:0
marital:2
marital:3
marital:4
marital:5
marital:6
marital:7
resi_type:0.0
resi_type:1.0
resi_type:2.0
resi_type:3.0
resi_type:4.0
resi_type:5.0
phone:N
visa:0
visa:1
diners:0
diners:1
american:0
american:1
occ_type:0.0
occ_type:1.0
occ_type:2.0
occ_type:3.0
occ_type:4.0
occ_type:5.0
educational:0.0
educational:1.0
educational:2.0
educational:3.0
educational:4.0
educational:5.0
MONTHS_IN_RESIDENCE: <5
MONTHS_IN_RESIDENCE: 5-12
MONTHS_IN_RESIDENCE: 13-24
MONTHS_IN_RESIDENCE: 25-40
MONTHS_IN_RESIDENCE: 41-44
MONTHS_IN_RESIDENCE: 45-48
MONTHS_IN_RESIDENCE: 49-60
PERSONAL_MONTHLY_INCOME: <256.479
PERSONAL_MONTHLY_INCOME: 256-845.915
PERSONAL_MONTHLY_INCOME: 845-1238.872
PERSONAL_MONTHLY_INCOME: 1238-2221.265
PERSONAL_MONTHLY_INCOME: 2221-2515.983
PERSONAL_MONTHLY_INCOME: 2515-3011.18
PERSONAL_MONTHLY_INCOME: 3007-4485
PERSONAL_MONTHLY_INCOME: 4485-8911
PERSONAL_MONTHLY_INCOME: 8911-10900
PERSONAL_MONTHLY_INCOME: >10900
MONTHS_IN_THE_JOB: >1

Table A.2 Non Imbalanced Variables

Features Name	Coefficients	P Values
AGE: <16.9	-0.589677	0.67925747

AGE: 16.9-30	0.194936	0.10003358
AGE: 31-43.3	0.437502	0.00008785
AGE: 43-56.5	0.678415	0.00000000
AGE: 56-66.4	0.940607	0.00000000
AGE: 66-82.9	0.929220	0.00000000
company:N	-0.037417	0.18497463
marital:0	0.120185	0.52251797
marital:2	0.230831	0.00000000
marital:3	0.334845	0.00368338
marital:4	0.150715	0.00382821
marital:5	0.006270	0.93309263
marital:6	0.004700	0.93779071
marital:7	0.065555	0.56878627
master:1	0.216545	0.00000011
MONTHS_IN_RESIDENCE: <5	-0.082853	0.09447630
MONTHS_IN_RESIDENCE: 13-24	-0.078230	0.14286724
MONTHS_IN_RESIDENCE: 25-40	-0.126421	0.05241553
MONTHS_IN_RESIDENCE: 41-44	0.120821	0.65291699
MONTHS_IN_RESIDENCE: 45-48	-0.428888	0.02963118
MONTHS_IN_RESIDENCE: 49-60	0.454442	0.02498902
MONTHS_IN_RESIDENCE: 5-12	-0.178711	0.00025607
MONTHS_IN_THE_JOB: >1	0.244749	0.46839155
occ_type:0.0	0.097355	0.11334576
occ_type:1.0	0.121795	0.00800801
occ_type:2.0	0.082036	0.03303837
occ_type:3.0	0.307599	0.04440738
occ_type:5.0	0.003271	0.93786093
payment_day:1	0.206962	0.00223104
payment_day:10	0.354666	0.00000000
payment_day:15	0.222431	0.00000002
payment_day:20	0.131151	0.00428280
payment_day:5	0.375113	0.00000000
payment_day:-99999	0.679516	0.01959612
PERSONAL_MONTHLY_INCOME: <256.479	0.065208	0.36141348
PERSONAL_MONTHLY_INCOME: >10900	-0.076773	0.78090261
PERSONAL_MONTHLY_INCOME: 1238-2221.265	0.032527	0.52590830
PERSONAL_MONTHLY_INCOME: 2221-2515.983	0.021316	0.87578178
PERSONAL_MONTHLY_INCOME: 2515-3011.18	0.450863	0.00411968
PERSONAL_MONTHLY_INCOME: 256-845.915	0.101183	0.00127533

PERSONAL_MONTHLY_INCOME: 3007-4485	0.619195	0.00064313
PERSONAL_MONTHLY_INCOME: 4485-8911	0.060248	0.75240713
PERSONAL_MONTHLY_INCOME: 8911-10900	1.069.960	0.21840332
product:1	0.218018	0.00118279
product:2	0.252568	0.00071306
prof_code_12_11	0.128503	0.01897827
prof_code_13_16_6_17_5_18	0.210412	0.03635009
prof_code_2_10	-0.072526	0.16919379
prof_code_4_15_3	-0.404729	0.10349845
prof_code_9	-0.019267	0.60719917
prof_code_7_8	-0.383192	0.00410314
resi_type_0_5_4	0.102109	0.07944647
resi_type_1	0.173226	0.00000690
resi_type_3	0.533208	0.02654136
state_resi:AC_DF_AM_AL	0.170607	0.15146891
state_resi:BA_PE	0.302444	0.00719020
state_resi:MS_MT_RJ_MA_CE	0.336979	0.00280343
state_resi:PA_PI_MG	0.420637	0.00026043
state_resi:RN_ES_GO_RR	0.287764	0.01318133
state_resi:SC_RO	0.798380	0.00000002
state_resi:SP	0.399935	0.00039836
state_resi:TO_AP_PR_PB_RS	0.539673	0.00000200
educational:1.0	0.611724	0.15355550
educational:2.0	-0.055117	0.70778048
educational:3.0	0.232942	0.04304150
educational:4.0	0.228840	0.04179624
educational:5.0	1.026.549	0.09435274
educational:6.0	0.091741	0.00088754