Executive Summary (400 Words)

This document reports solutions to the following two named problems: 'Age' and 'Churning Scrutiny'. A logistic regression classification model was designed for the first problem, the aim of which was to determine the ability of a customer's age to determine whether they may be eligible for a credit card. Secondary to this was to determine the ability of two other features to do the same. Data cleaning and a Box Tidwell test were carried out to determine suitability of the age feature for a logistic regression. The data were split into training and testing subsets and the logistic regression model was trained and tested using these data, revealing very little information about the ability of age to determine credit card ownership. A confusion matrix and accuracy score of 53% illustrate that age is not a good predictor of credit card ownership.

Other features were selected as predictors for credit card eligibility using Chi squared and Kruskal-Wallis tests, point-biserial correlations and variance inflation factor calculation. Two logistic regression models were constructed using tenure and Own_property as predictor variables. Similar to the model trained on age, the accuracies of these models were 51% and 58%, respectively. The area under a ROC-curve was used as a metric to determine how close to random these classification models were. For tenure and Own_property, these values indicate better than random classification of the credit_card class.

The second problem was purely statistical in nature. Spearman's ρ correlation coefficients were used to determine an association between balance and tenure attributes. The result of this shows very little association, potentially owing to the interval nature of the tenure attribute. Next, a similar process was carried out for all other attributes in the data, using point-biserial correlation, Spearman's ρ and Kruskal-Wallis tests to determine associations between other attributes and balance. Of all of these, eight attributes were found to have significant associations with balance under specific circumstances.

In repeated study, it may be useful to test other classification models to see if they provide an increase in classification performance over logistic regression. A naïve Bayes classifier or support vector machine could be good alternative candidates, or a simple decision tree classifier. For the Churning Scrutiny problem, other machine learning solutions could have been possible that may have provided me with more tangible results than what was calculated here using statistics.

Task 1: Discussion of Techniques

Problem 1: 'Age'

The question of whether age is a good predictor of credit card eligibility is a classification problem – can we use the age attribute of this dataset to determine whether a customer is eligible for a credit card?

Logistic Regression

To answer this, a suitable classification method needs to be selected. The logistic regression classification model is effective at classifying binary and multinomial data using multiple predictor variables. This type of regression mandates that four conditions be met [1]: independent measurements; linearity between variables and the logarithmically transformed response variable; absence of multicollinearity between variables; and sufficient sample size. In the case of univariate regression, there is no multicollinearity, so this can be disregarded. Given that each instance in the dataset is an individual customer, it can be assumed that each is an independent measurement, and linearity between the log of the response variable and continuous variables can be tested using the Box-Tidwell approach [2].

The benefit of using logistic regression is that it places few restrictions on the data – the predictor variable can be continuous, ordinal, or nominal provided dummy variables are used. This allows us to examine each of the predictor variables using univariate and multivariate analysis, regardless of the mixture of data types. Furthermore, this does not require the data to exhibit any probability distribution. The only data type restriction is that the outcome variable must be categorical.

Logistic regression starts to falter as complexity increases – this is because the loss function (Eq. 1) does not punish complexity, prohibiting a trade-off between performance and complexity as other classification models do. As a mathematical model, this method of classification struggles with missing data – it requires complete data to estimate a coefficient for each feature and it has no inherent imputation method.

$$\ln\left(\frac{\hat{y}}{1-\hat{y}}\right) = A + \sum B_j X_{ij} \qquad \text{Eq. 1}$$

Data Cleaning

Before modelling our data, we need to check that our age feature meets all requirements of the logistic regression model. The Box-Tidwell transformation [2] provides a way of testing for linearity between the log predictor and outcome. In performing a logistic regression using the log of the predictor, we can use a significant p-value to indicate a violation of the assumption – a value smaller than 0.05 means the log of the predictor is linearly proportional to the outcome. Testing age with the Box-Tidwell method indicates that this linearity assumption is not violated (p=0.807).

The provided dataset does not have any missing values, so no imputation is required. What remains to be reviewed is the sample size. Studies show that sample size for meaningful logistic regression should be at least $100 + 50i$ where i is the number of independent variables [3], [4]. This dataset has 22 features and 9709 entries, so the minimum sample size of 1200 is far exceeded.

Age as a Predictor

For most banks (Halifax, HSBC, Lloyds, Barclays, Natwest), the minimum age to apply for a credit card is eighteen. Explicitly, this is the only relationship between age and credit card eligibility. Classification models can be used to see if an implicit relationship may exist between the two.

The provided dataset includes no customers under the age of eighteen, so we can discount the influence of under-18s being refused a credit card. The data were split into training and test subsets and the training set was used to train a logistic regression model using class weight balancing. This weight balancing parameter is required by the imbalance in the credit_card class – the ratio of customers with credit cards versus those without is roughly 7 to 3.

Despite the rationale behind age being a predictor of credit card ownership, the logistic regression model had a hard time identifying any tangible relationship. The accuracy of the model is estimated using Eq. 2 and is visualised in Fig. 1 (Appendix) using a confusion matrix. The resultant accuracy for the model trained on age is 53%, meaning this model does little better than randomly assigning class attributes based on the predictor.

$$accuracy(y,\, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \mathbb{I}(\hat{y_i} = y_i) \tag{Eq.2}$$

Furthermore, the coefficient associated with age is very small (-0.001) and the p-value for this is not significant (0.345 > 0.05), which tells us that there is little linear association between age and credit card ownership and should not be used as a reliable predictor of credit card ownership.

Other Features as Predictors

In deciding which other features could be used as univariate predictors of credit card ownership, there are two main methods two approach this: research and statistical inference. In this section, I use statistical inference to choose two features that can be used as predictors of the credit_card feature.

Statistical Inference

From a statistical point of view, there are a few methods we can use to isolate some features that may be good predictors of credit card ownership. Primarily, correlation coefficients are effective at showing statistical significance of a relationship between explanatory and response variables. Depending on the data type and probability distribution of a feature, the choice of correlation coefficient is crucial. Point-Biserial correlation is chosen to find the correlation between a continuous variable and the binary response. A p-value associated with this statistic can be used to tell us if the correlation is significant. To study the association between other data types, a χ^2 test is used to evaluate association for dichotomous features; for multinomial features, a non-parametric Kruskal-Wallis test. The associated p-values with these can be used to see if any correlation between a feature and credit_card is significant and warrants further investigation.

From this analysis, two features had a significant association – home ownership and tenure. Given that home ownership is linked to credit score by mortgages, this association with credit card eligibility is logical.

We can also decide which features to use based on which ones contribute the most to a logistic regression model that uses all features. To do this, combat the presence of multicollinearity. Multicollinearity is a property of features in a dataset that have a correlation with each other, and it can be evaluated using a variance inflation factor (VIF). Typically, values above five show high collinearity with another feature and should be either dropped or transformed to reduce this inflation. Interestingly, the features that have a high VIF value are age and credit score. As such, they are dropped from the multivariate logistic regression.

Now, to select features that could be used as predictors of credit card ownership, the remaining features are ranked by their coefficients in a multiple logistic regression. The higher the absolute value of the coefficient, the more relative importance they have. From this, we can isolate tenure as a predictor.

Using selection from both coefficient and association analysis, we have two suitable candidates: **tenure** and **Own_property**.

Performance of tenure and Own_property

Performing logistic regression using tenure as the exogenous variable, we achieve an accuracy of 50%, and for Own_property, the accuracy is 58%. Neither of these are particularly large and could indicate spurious classification. Both features were tested using an ROC to determine how close to total randomness each learned model is. The area under curve (AUC) tells us how closely our model performs to chance. A value of 1 shows perfect modelling, while 0.5 shows random modelling. For the model with tenure, the AUC-ROC is 0.529, and that for Own_property is 0.511. Again, these are not good values, but they tell us that modelling using these features is not entirely spurious.

Problem 2: 'Churning Scrutiny'

In this problem, the bank wants to know if there is an association between balance and tenure. To answer this question, a statistical approach was used.

Spearman's ρ Analysis of Balance and Tenure

To determine the presence of a link between balance and tenure, I used Spearman's ρ correlation coefficient [5]. This is a non-parametric analogue of Pearson's R correlation coefficient that indicates the strength, direction and significance of correlation between two features that are not drawn from any probability distribution. In this case, tenure does not fit into any common distributions, as tested using the Kolmogorov-Smirnov test. However, balance does exhibit normality under certain conditions, notwithstanding that it doesn't affect this portion of the analysis. Normality is tested for using the D'Agostino normality test [6].

Spearman's ρ analysis of the relationship between balance and tenure revealed very little – statistic = -0.009, p=0.358. The correlation is very weak, and it is not statistically significant. Surprisingly, if we remove instances where $balance = 0$ and $Exited = 1$, the distribution of balance becomes closer to normal. Typically, we'd expect the performance of a correlation coefficient to increase if a feature is parametric. However, in this instance, the statistic becomes 0.001 and the p-value becomes 0.939 (hugely insignificant).

Correlation Analysis of Other Attributes

The second aim of this solution is to find other attributes that may have a correlation with balance, and particularly those of high value. To this end, a few masks for the balance were created to filter out the following: $balance = 0$; $balance = 0, Exited = 1$; $balance < quantile_{95}(balance)$; $balance < quantile_{99}(balance)$.

For each of the features, the correlation coefficient was chosen based on the data type. Dichotomous features used Point-Biserial correlation; continuous, Spearman's ρ; multinomial features used a Kruskal-Wallis test with a Conover-Iman post-hoc test to identify significantly different groups. These are all non-parametric indicators of association. Again, the statistic of the correlations (Spearman and Point-Biserial) indicates strength and direction, while the p-value of each method indicates significance. Each feature was filtered using each of the masks above and tested for association with balance. Of these analyses, eight features had a statistically significant association: Exited, gender, age, Own_car, Geography, Total_income, credit_card, and Housing_type. Association between four of these is shown in Fig. 2.

Of these, the strongest correlations are Total_income (statistic = 0.21), credit_card (statistic = -0.22) and Geography (p-value = 2.05×10^{-294}). From this analysis, it can be concluded that a customer is more likely to be a higher earner if they are from the Netherlands, if they *don't* have a credit card, or if they are high earners (in the 99th quantile of balances).

Task 2: Evaluation of Tools and Languages

The analysis of these data was carried out using Python version 3.12 in an Anaconda-managed environment running JupyterLab [7], [8], [9]. Python is an open-source programming language with about 8.2 million users. Because it is open-source, there are many contributors developing useful packages, many of which are useful for data mining and statistical analysis. In particular, the scikit-learn, scipy and statsmodels packages were particularly useful as they provide a flexible interface with some otherwise complex programming concepts. Anaconda is a useful environment management tool that allows me to organise Python packages, as well as launch JupyterLab. Jupyter Notebooks are useful tools for writing notes on my code and improving legibility and maintainability.

Despite its flexibility, Python comes with a learning curve since it is a programming language. Processes are not abstracted behind a GUI like with other tools like WEKA and a deeper understanding of these processes may be required to choose the most suitable packages for a given scenario (e.g. statsmodels provides statistical insights to data mining models, while scikit-learn simply generates a model for use in machine learning).

If I were to complete this task again, I would likely use R instead of Python. R is also an open-source programming language with many contributors, but it is designed specifically with data analysis in mind (compared with the all-purpose nature of Python). Because of this, it lends itself well to statistical models like logistic regression and integrates well with association analysis techniques. The main benefit of R over Python is that R data structures are homogenised – using data.frame as its main structure, rather than a choice of the native list type, numpy.ndarray, pandas.DataFrame, or even polars.DataFrame to confuse things [10]. This makes it an easier choice for learning statistics and data mining techniques and is also compatible with Anaconda development environments and Jupyter Notebooks.

On the other hand, Python is more industry-standard than R is, at least outside of academia. Python is integrated into significantly more business tech stacks than R is (260,000 companies use Python vs 6600 using RStudio [11]). Python also lends itself to integration with other use cases like web app development. Machine learning using scikit-learn can be integrated into a Django webapp, or data can be obtained using web scraping packages and cleaned for use in statsmodels scripts. This extensibility isn’t present in R.

Task 3: Discussion of Current Literature

Current Issues in Data Classification in Credit Risk

Data Classification is an old concept, but with the age of big data and ever-changing rules, regulations and applications, there are always avenues for further development and solutions [12].

Classification Model Transparency

Typically, classification models like naïve Bayes, artificial neural networks and support vector machine models are discriminative – they don't provide any information about *how* they make classification predictions; they only provide predictions. Some of the most powerful machine learning models used in credit risk assessment are discriminative models – a proportion that is likely to increase with the advent of artificial intelligence [13]. The Basel II Accord [14] is a framework that increases regulations on credit scoring agencies that requires them to be transparent with how credit scores are determined. This transparency is difficult to obtain when using deterministic classification models to formulate a credit risk report, and the alternative is to use poor-performing generative models like linear regression and decision trees [13].

Under European GDPR regulations, any customer is also entitled 'not to be subject to a decision based solely on automated processing' [15], implying that any credit scoring must be transparent on the instance level, as well as the dataset level. As such, questions that must be addressed include 'Is the model well fitted around this prediction?', 'Which variables contribute to the selected prediction?', and 'What is the model prediction for this instance?' [16].

Various methods have been developed to adhere to these rules and regulations. One good contribution to the literature is the Transparency, Auditability, and eXplainability for Credit Scoring (TAX4CS) framework (Fig. 3) [16]. This model is used to ensure that all stakeholders are considered when using discriminative classification models, and that appropriate techniques are used to convey how a particular prediction was made, as well as evaluate the entire model. Section 4 of Bücker et. al's work outlines specific metrics that can be used to determine how an instance's prediction was made. Primarily, these are SHAP and iBreakDown [17], [18], which are recent additive and non-additive local model prediction interpretation frameworks, respectively. iBreakDown is a powerful tool for identifying how the value of each attribute used in training a model influences the decision-making process of that model on an instance level (Fig. 4).

SHAP (Shapley Additive exPlanations) is a model-agnostic method that derives from game theoretical Shapley values. These values are additive, and they tell us how the value of an attribute affects the model outcome prediction. The additivity of the SHAP method is advantageous as it can be used to illustrate visually how each attribute contributes to the final prediction value. However, this additivity is an issue when there are interactions between features because these interactions imply an order to which features should be 'read'. These interactions are not captured by additive model-agnostic methods like SHAP and LIME. iBreakDown is a non-additive model-agnostic that incorporates interactions in the explanations.

iBreakDown has seen usage in several recent papers [19], [20], [21] that show that this algorithm is effective at improving the interpretability of so-called black box machine learning models in applications both within and outside of finance. I believe that the TAX4CS framework in conjunction with model-agnostic tools like iBreakDown will prove to be effective. The creators of iBreakDown, however, have their own framework similar to TAX4CS named DALEX [22]. This also acts as a Python software suite that unifies popular machine learning suites (like sklearn and tensorflow) and integrates iBreakDown, as well as SHAP and other explainer methods for use alongside them. There is also an R port for this that provides similar capabilities.

With respect to this frontier, I believe that use of a framework promoting machine learning transparency like TAX4CS or DALEX provides a working solution to the problem posed by regulations such as EU GDPR and Basel II.

Appendices

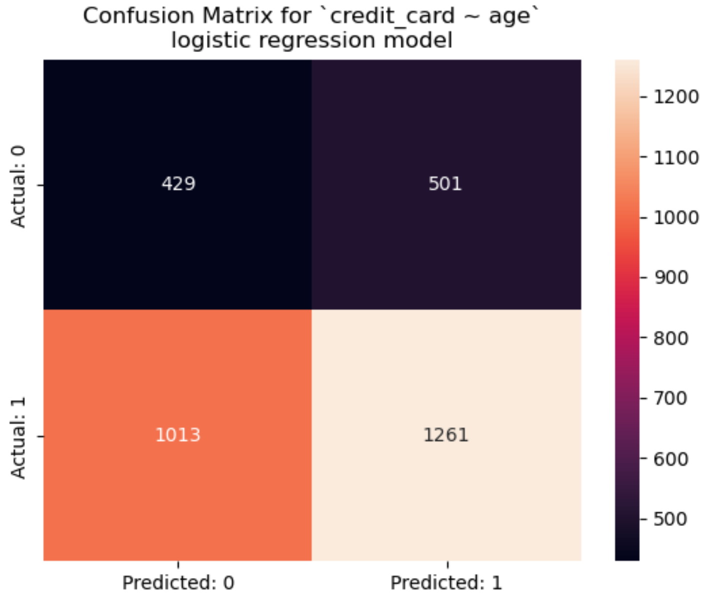

Figure 1 – Confusion matrix for evaluating the effectiveness of logistic regression using age as a predictor of credit card ownership.

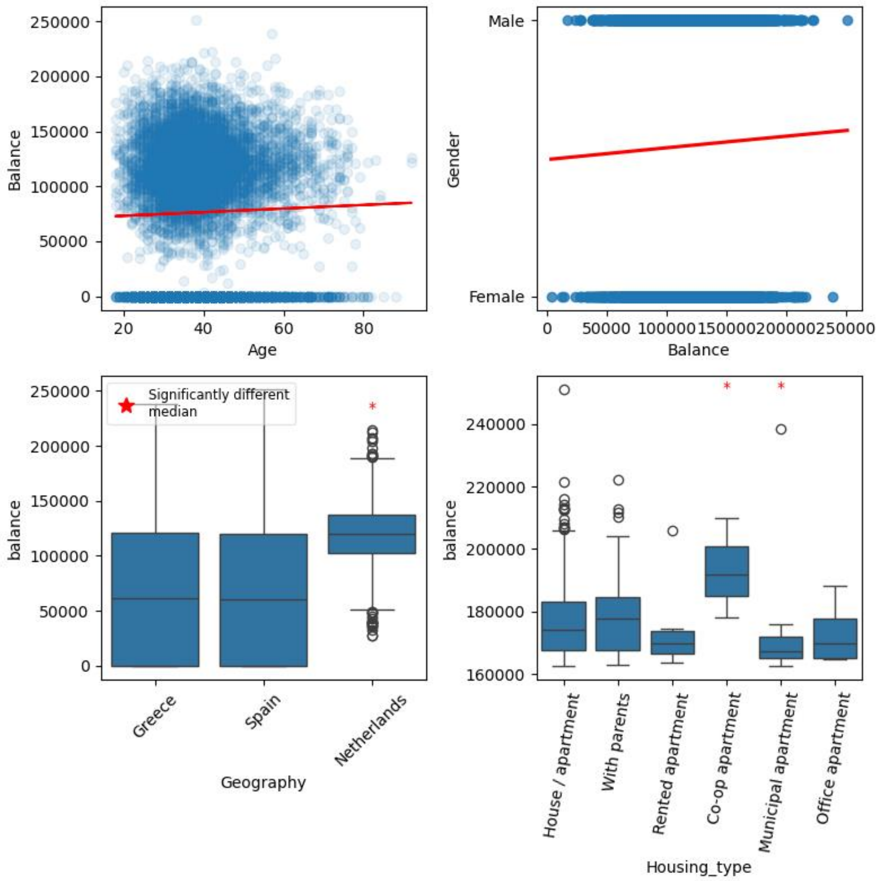

Figure 2 – Associations between balance and four other features – Age, Gender, Geography and Housing Type

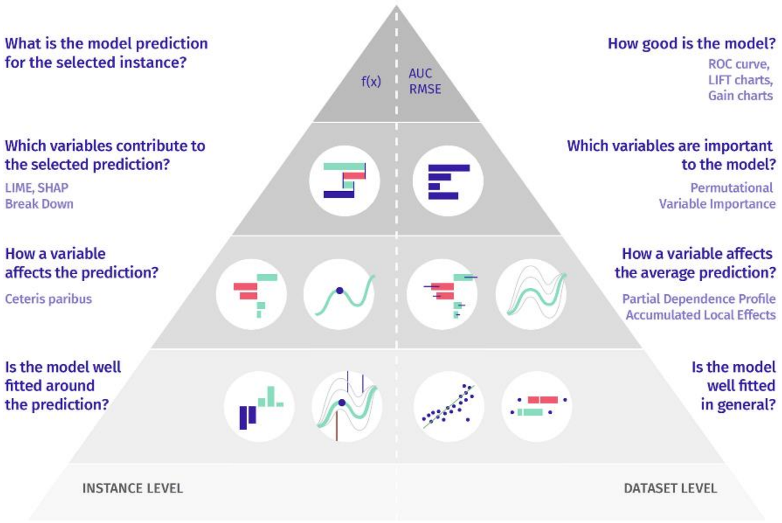
What is the model prediction for the selected instance?
f(x)
AUC
RMSE
How good is the model?
ROC curve,
LIFT charts,
Gain charts
Which variables contribute to the selected prediction?
LIME, SHAP
Break Down
Which variables are important to the model?
Permutational
Variable Importance
How a variable affects the prediction?
Ceteris paribus
How a variable affects the average prediction?
Partial Dependence Profile
Accumulated Local Effects
Is the model well fitted around the prediction?
Is the model well fitted in general?
INSTANCE LEVEL
DATASET LEVEL

Figure 3 – TAX4CS framework laid out by Bücker et al. (2022) aims to introduce transparency throughout the discriminative modelling process

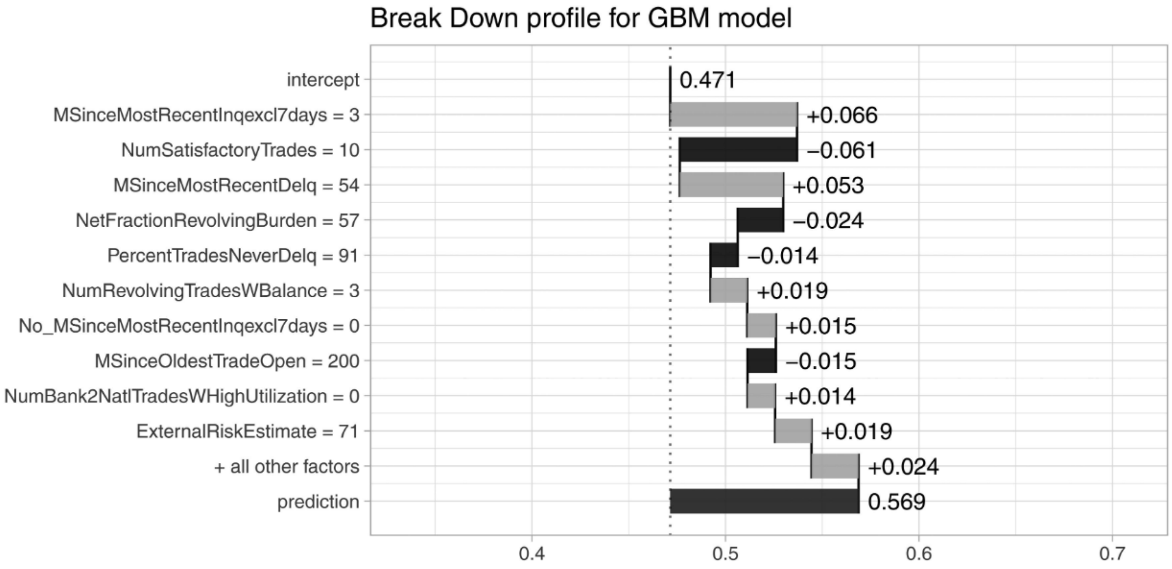

Figure 4 – from [16], iBreakDown profile for a Gradient Boosting classification model indicating which attributes have the most influence on the final prediction value. We can see that MSinceMostRecentInqexcl7days has the most influence on the final outcome, which can be reported by credit scorers.

References

[1] B. G. Tabachnick, L. S. Fidell, and J. B. Ullman, *Using multivariate statistics*, vol. 6. pearson Boston, MA, 2013.

[2] G. E. P. Box and P. W. Tidwell, ‘Transformation of the Independent Variables’, *Technometrics*, vol. 4, no. 4, pp. 531–550, Nov. 1962, doi: 10.1080/00401706.1962.10490038.

[3] M. A. Bujang, N. Sa’at, T. M. I. Tg Abu Bakar Sidik, and L. Chien Joo, ‘Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data’, *Malays. J. Med. Sci.*, vol. 25, no. 4, pp. 122–130, 2018, doi: 10.21315/mjms2018.25.4.12.

[4] C. C. Serdar, M. Cihan, D. Yücel, and M. A. Serdar, ‘Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies’, *Biochem. Medica*, vol. 31, no. 1, p. 010502, Feb. 2021, doi: 10.11613/BM.2021.010502.

[5] C. Spearman, *The Proof and Measurement of Association Between Two Things*. in Studies in individual differences: The search for intelligence. East Norwalk, CT, US: Appleton-Century-Crofts, 1961, p. 58. doi: 10.1037/11491-005.

[6] R. D’AGOSTINO and E. S. PEARSON, ‘Tests for departure from normality. Empirical results for the distributions of b2 and √b1’, *Biometrika*, vol. 60, no. 3, pp. 613–622, Dec. 1973, doi: 10.1093/biomet/60.3.613.

[7] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[8] ‘Anaconda Software Distribution’, *Anaconda Documentation*. Anaconda Inc., 2020. [Online]. Available: https://docs.anaconda.com/

[9] T. Kluyver *et al.*, ‘Jupyter Notebooks – a publishing format for reproducible computational workflows’, in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, F. Loizides and B. Schmidt, Eds., IOS Press, 2016, pp. 87–90.

[10] C. Ozgur, T. Colliau, G. Rogers, and Z. Hughes, ‘MatLab vs. Python vs. R’, *J. Data Sci.*, vol. 15, no. 3, pp. 355–372, Mar. 2021, doi: 10.6339/JDS.201707_15(3).0001.

[11] ‘Python commands 3.21% market share in Programming Languages’. Accessed: Oct. 27, 2024. [Online]. Available: https://enlyft.com/tech/products/python

[12] X. Zhang and L. Yu, ‘Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods’, *Expert Syst. Appl.*, vol. 237, p. 121484, Mar. 2024, doi: 10.1016/j.eswa.2023.121484.

[13] B. R. Gunnarsson, S. vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, ‘Deep learning for credit scoring: Do or don’t?’, *Eur. J. Oper. Res.*, vol. 295, no. 1, pp. 292–305, Nov. 2021, doi: 10.1016/j.ejor.2021.03.006.

[14] ‘Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework’, Jun. 2004, Accessed: Oct. 27, 2024. [Online]. Available: https://www.bis.org/publ/bcbs107.htm

[15] ‘General Data Protection Regulation (GDPR) – Legal Text’, General Data Protection Regulation (GDPR). Accessed: Jun. 24, 2024. [Online]. Available: https://gdpr-info.eu/

[16] M. Bücker, G. Szepannek, A. Gosiewska, and P. Biecek, ‘Transparency, auditability, and explainability of machine learning models in credit scoring’, *J. Oper. Res. Soc.*, vol. 73, no. 1, pp. 70–90, Jan. 2022, doi: 10.1080/01605682.2021.1922098.

[17] S. Lundberg and S.-I. Lee, ‘A Unified Approach to Interpreting Model Predictions’, Nov. 25, 2017, *arXiv*: arXiv:1705.07874. doi: 10.48550/arXiv.1705.07874.

[18] A. Gosiewska and P. Biecek, ‘Do Not Trust Additive Explanations’, May 08, 2020, *arXiv*: arXiv:1903.11420. doi: 10.48550/arXiv.1903.11420.

[19] Y. Xu, W. Zhang, X. Ma, M. Wu, and X. Jiang, ‘Retrospective analysis of interpretable machine learning in predicting ICU thrombocytopenia in geriatric ICU patients’, *Sci. Rep.*, vol. 14, no. 1, p. 16738, Jul. 2024, doi: 10.1038/s41598-024-67785-1.

[20] Y. Zhang, L. Zhang, H. Lv, and G. Zhang, ‘Ensemble machine learning prediction of hyperuricemia based on a prospective health checkup population’, *Front. Physiol.*, vol. 15, Apr. 2024, doi: 10.3389/fphys.2024.1357404.

[21] V. D’Amato, R. D’Ecclesia, and S. Levantesi, ‘Firms’ profitability and ESG score: A machine learning approach’, *Appl. Stoch. Models Bus. Ind.*, vol. 40, no. 2, pp. 243–261, 2024, doi: 10.1002/asmb.2758.

[22] H. Baniecki, W. Kretowicz, P. Piątyszek, J. Wiśniewski, and P. Biecek, ‘dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python’, *J. Mach. Learn. Res.*, vol. 22, no. 214, pp. 1–7, 2021.