German credit risk classification using Machine Learning Algorithms

Abstract—**This paper provides and discusses the results of applying three different Machine Learning classification techniques to German Credit Data to decide whether to issue a loan approval or not. The goal of the German Credit Data classification is to increase the likelihood of earning from profitable loans while minimizing the likelihood of making risky loans to applicants. We go on to explore other issues such as data imbalance, model transparency, and bias. The applicant's socioeconomic and demographic profiles are taken into account before making a decision about his or her loan application.**

Keywords—loan approval; German Credit data; machine learning

I. INTRODUCTION

Credit and trust are the pillars of contemporary financial systems. A key factor that gauges and forecasts a debtor's chances of defaulting is credit risk. For the entire system, accurate credit risk estimation is crucial. Breakdowns in the calculation of credit risk can result in systemic failures, like the subprime crisis of 2008 [1].

The most important initial stage in the development of credit rating methodologies is the study of Durand in 1941 [2]. Risk variables were identified, and for the first time, their weights were computed empirically. Even so, default risk-related behavioural traits and historical payment history were not taken into account [3]. Credit scoring methods at the time neglected applicant qualities that are directly and intricately tied to believability [4]. However, when computers and statistical programmes became more widely available and potent after the 1980s, the emphasis shifted to finding new predictor factors [5]. The findings of Onay and ztürk (2018) indicate that "statistical methodologies and accuracy" and "new determinants of creditworthiness" are the most researched areas in credit risk-related research [6].

Since 2010, there has been a surge in the number of studies looking at fresh creditworthiness-related factors. Additionally, the regulations updated to address the challenges of risk management in the digital era have increased the output of pertinent research in the area [6]. Prior to the development of automated credit scoring systems, decisions were made using subjective criteria. Important disadvantages of this strategy are its inability to process a huge volume of applications and its susceptibility to misclassification mistakes and inaccuracies. By using machine learning methods, this problem is undoubtedly resolved but numerous ethical issues are raised by the usage of algorithms. One concern is how the credit scores can misclassify applicants and lead to stigmatization. Additionally, because of issues with the General Data Protection Regulation (GDPR) directive, we must exercise caution while using personal data [7]. Decision-making and transparency depend on the ability to estimate credit risk, monitor it, trust models, and handle loans efficiently.

Machine learning methods like Gradient boost, Random forest (tree-based classifiers), and support-vector machines are particularly well suited for consumer credit because of the high sample sizes and the intricacy of the potential correlations between consumer transactions and characteristics [8]. This study focuses on these algorithms to determine the choices while taking the applicant's socioeconomic and demographic profile data provided by German Credit Dataset into account. The data set is covered in more detail in part II. The machine learning classification methods that were used on the data set are briefly reviewed in section III. The procedures needed to pre-process the data are covered in section IV. The outcomes of using machine learning algorithms on the data set are shown and discussed in sections V and VI. Link to the Jupyter Notebook is provided in the appendix and can be used to replicate the results precisely.

II. THE DATA SET

The German Credit Dataset is provided by Prof. Hofmann and is publicly accessible from the UCI Machine Learning Repository [9]. One thousand loan applicants are divided into two categories: : good credit risks and bad credit risks, based on 20 criteria. There are no missing values in this dataset.

TABLE I

ATTRIBUTES OF THE DATASET

Attribute	Description
checking_acc	Status of existing checking account
duration_month	Duration of loan in months
credit_hist	Credit history of the applicant
purpose	Purpose of the loan
credit_amount	Credit amount
savings	Savings account/bonds
employment_history	Present employment since (years)
installment_rate	Installment rate in percentage of disposable income
sex	Male/Female
loan_status	Other debtors / guarantors for the applicant
resident_since	Present residence since
property	Property type of applicant
age	Age in years
installment_plan	Other installment plans
housing_type	Rent, own or for free
existing_cards	Number of existing credits at this bank
job	unskilled nonresident, unskilled resident, skilled or highly skilled
liability_count	Number of people being liable to provide maintenance for
telephone	Is the Telephone registered or not - yes/no
foreign_worker	Is the applicant a foreign worker – yes/no
credit_score	Whether the issued loan was a good decision or bad

Since the dataset takes into account demographic and socioeconomic factors, it may lead to a biased model in different ways. Age and sex, for instance, can contribute to the creation of a skewed model. As seen in Figure 1, the dataset represents more men than women. Additionally, characteristics like income might distinguish between groups of people who are privileged and those who are not.

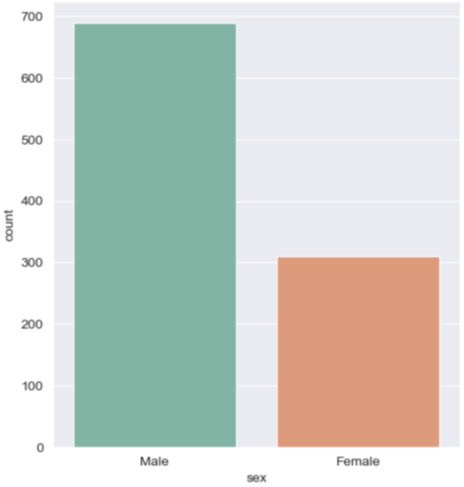

Fig. 1. Male/Female ratio in the dataset

In order to find out the most affecting features in the dataset, the model needs to be tested with feature selection algorithm. Bias is further reduced by hyperparameter tuning, and the outcomes are cross-validated.

The bank's choice is subject to two different kinds of risks:

1) If the applicant has good credit risk, which means they are likely to repay the loan, then the bank loses business by turning them down.

2) If the applicant has a poor credit risk, meaning they are unlikely to pay back the loan, the bank will suffer a loss by issuing the loan.

III. CLASSIFICATION TECHNIQUES

A. Random Forest

In a nutshell, Random Forest is a collection of Decision Trees and is an extension of bagging ensemble method[10]. With more trees, the model is more effective, which increases model accuracy. Random Forests operate by building a large number of decision trees during the training phase, with the output (class) being the average of all the classes. Random forest's inherent capacity to adjust for decision trees tendency to overfit their training set is its most practical advantage. When applying this technique, the overfitting issue is almost entirely eliminated by the bagging method and random feature selection. Additionally, Random Forest typically retains its accuracy even when some data is absent [10]. The tuned parameters for Random Forest are the number of trees, the maximum depth of each tree, the number of features to consider when looking for the best split, and the function to gauge the split's quality.

B. *Gradient Boosting Classifier*

Gradient Boosting is an ensemble method. In gradient boosting, each prediction seeks to outperform the one that came before it by reducing the inaccuracies. Instead of fitting a prediction to the data at each iteration, Gradient Boosting fits a new predictor to the residual errors produced by the previous predictor [11]. The tuned parameters for this classifier are the learning rate, number of boosting stages to perform, the fraction of samples to be used for fitting the individual base learners and the maximum number of nodes in the tree.

C. *Support Vector Machine (SVM)*

SVM carries out classification by determining a hyperplane or line for a two-dimensional example, or a maximum margin of separation between the classes. For non-linear separable models, a kernel function can be used, and the original input vectors are transferred to a higher dimensional feature space. The goal is to maximize the hyperplane margin while obtaining the fewest possible misclassifications. Gamma and C (Cost) are the parameters that are tuned for higher accuracy. The distance between the nearest sample and the hyperplane is known as the margin gamma. The model's decision to prioritize margin maximization or reducing classification error is controlled by the penalty parameter C (cost).

IV. EXPERIMENTAL SETUP

This part covers data preparation, obtaining the key model parameters, testing the model by predicting the test set, and, finally, estimating the model's performance and tweaking the hyper-parameters to enhance the model.

A. Data Preprocessing :

1) Data Imputation: Since the dataset didn't contain any Nan values or inconsistent values, no data replacement was necessary.

2) Categorical feature encoding: Since math is typically done using numbers, categorical data is encoded using numbers. If the data is not numerical, machine learning algorithms cannot run and process it. The 13 categorical features have been encoded using OneHotEncoder function to convert the features into binary values. Further, the dummy variables are eliminated after encoding.

3) Data Standardization: StandardScaler from python sklearn library was used for standardization. StandardScaler follows Standard Normal Distribution (SND). As a result, it adjusts the data to unit variance and sets mean = 0. Not all machine learning models need standardization.

Out of the 3 models used in this paper, Random Forest and Gradient boosting classifier do not require the data to be standardized because they are not sensitive to the magnitude of the variables. However, standardization doesn't affect these models either way. The third model used in this paper is Support Vector Machine (SVM), and in SVM optimization happens by minimizing the decision vector, and the ideal hyperplane is affected by the scale of the input features, so it is advised to standardize the data for the SVM model [12]. Hence, the dataset in this study has been standardized.

4) Class Imbalance: There are 700 labels for good credit scores and 300 labels for low credit scores in the dataset, and this imbalance would make the machine learning system perform poorly.

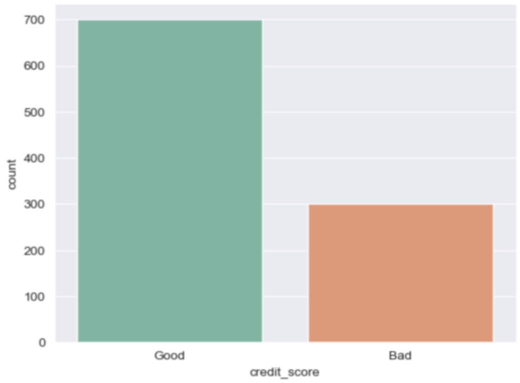

Fig. 2. Class Imbalance

To overcome this, three techniques were considered, but only two were adapted.

a) Random Over-sampling: Random over-sampling is performed by randomly duplicating instances of minority class to match the number of instances from majority class. In this paper, 300 minority instances with a bad credit score were selected and merged with 700 instances with a good credit score to create a new dataset of 1400 instances.

b) SMOTENC (SMOTENC for Numeric and Categorical data): It is an extension of SMOTE and can handle both numerical and categorical features. It works by first applying SMOTENC to the numerical features, and then using the nearest neighbors of the synthetic samples to determine the values for the categorical features. This helps oversample the minority class while maintaining the links between the features in a dataset containing both numerical and categorical features [13].

c) Random Under-sampling: This method is rejected for this paper as under-sampling works by deleting the majority class and can result in losing valuable information. Since the dataset uses personal information, losing data could make the model more prone to bias.

5) Dimensionality Reduction: PCA is a common technique used for dimensionality reduction and works by finding underlying correlation between the variables. As seen in figure 3, there are no strong correlation between the features in the dataset and hence PCA wasn't performed on the data.

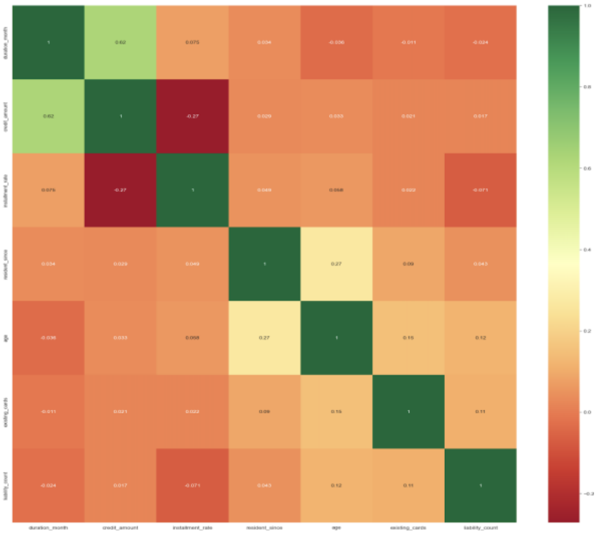

Fig. 3. Correlation Heat map

6) Feature Selection: After categorical feature encoding, the dataset comprises 45 features, hence feature selection was done on the data. In contrast to PCA, feature selection merely selects and excludes predetermined features without altering them. Two feature selection techniques were used based on the classifier model.

a) Recursive Feature elimination (RFE): In the Random Forest and Gradient Boosting classifiers, RFE was utilized, and it works by eliminating the weakest features from a model until the required number of features are reached. The features were chosen based on feature importance and reducing them down to 20 provided the models the best accuracy.

b) SelectKBest: When using SVMs with non-linear kernels, classification takes place in a space of features that has been altered rather than the original features. Therefore, it is impossible to rate the significance of unique traits and hence RFE cannot be used.

Hence, in order to determine the ANOVA F-value and assess the level of variance between the groups, SelectKBest is employed for SVC. The 20 attributes that produced the best results were chosen.

B. Model Selection: In this study, the performance scores of Random Forest, Gradient Boost, and SVM are compared. These models were chosen because they have been the subject of numerous studies and papers [2][14][15] examining how well they predict credit risk rating systems.

C. Hyperparameter Tuning: This helps to find the best set of parameters for each of the Machine Learning models. The results from the preceding iteration are used by the sequential model-based optimization process known as Bayesian optimization to select potential values for the following hyperparameters and this makes it faster than other techniques like random search and Grid search [16]. The ideal parameters as determined by BayesSearchCV are shown in Table II.

TABLE II BEST PARAMETERS

Over-sampled Data		
Random Forest	Criterion	Gini
	Max Features	Sqrt
	Max Depth	16
	No. of estimators	200
Gradient Boosting Classifier	Learning rate	0.01
	Max depth	10
	No. of estimators	250
	subsample	0.9
SVC	C	1.4
	Gamma	1000
	Kernel	rbf

SMOTENC		
Random Forest	Criterion	Gini
	Max Features	Log2
	Max Depth	15
	No. of estimators	200
Gradient Boosting Classifier	Learning rate	0.15
	Max depth	8
	No. of estimators	300
	subsample	0.9
SVC	C	0.5
	Gamma	54.55
	Kernel	linear

D. Train/Test: The data must be randomly split into a training set and a testing set after the preceding procedures. In this example, the dataset is split into variables for the characteristics and labels that will be trained and tested using the scikit-learn function train test split, which also shuffles the dataset. 75% of the data in this project are used for training, while the remaining 25% are utilized for testing

E. Cross-Validation: Once the model is trained and tested, it's accuracy is cross validated using the scikit-learn function cross val score.

V. EXPERIMENTAL RESULTS

To select the appropriate model for the credit risk classification, accuracy and F-measure must be considered. High accuracy indicates that the model is making correct predictions most of the time. However, F score (harmonic mean of Precision and Recall) must also be taken into account because it provides a more accurate indicator of erroneously classified observations than accuracy. The overall effectiveness of the various machine learning algorithms and the results are shown in Table III.

TABLE III ACCURACY AND F-MEASURE SUMMARY

Type of Measurement	**Over-sampled**	**SMOTENC**
Random Forest		
Accuracy	86.0%	80.0%
F-score	85.28%	80.11%
Cross -validation 10 Fold	87.91%	78.00%
Time (seconds)	299.16	309.40
Gradient Boosting Classifier		
Accuracy	85.14%	83.42%
F-score	84.24%	83.79%
Cross -validation 10 Fold	88.23%	81.19%
Time (seconds)	458.15	427.40
Support Vector Machine		
Accuracy	92.0%	76.0%
F-score	92.89%	74.54%
Cross -validation 10 Fold	92.33%	74.38%
Time (seconds)	50.81	47.97

As seen in Table III, the Random Forest and Gradient Boosting classifier consistently performs for both SMOTENC and over sampled data with just a slight decline in overall accuracy and F-measure for SMOTENC. SVM, however, experiences a sharp decline in performance for SMOTENC data. SVM is a sensitive classifier, and it may be influenced by the presence of noisy or incorrect samples in the dataset. If the synthetic samples generated by SMOTENC are incorrect or noisy, this can lead to decreased performance of the SVM model.

Table III and Fig.5 gives us a good insight of the overall performance of the models. A ROC curve helps evaluate the trade-off between the true positive rate and the false positive rate. Although table III shows the highest accuracy for SVM (over-sampled), the ROC-AUC curve shown in the first image of Fig.4 suggests otherwise.

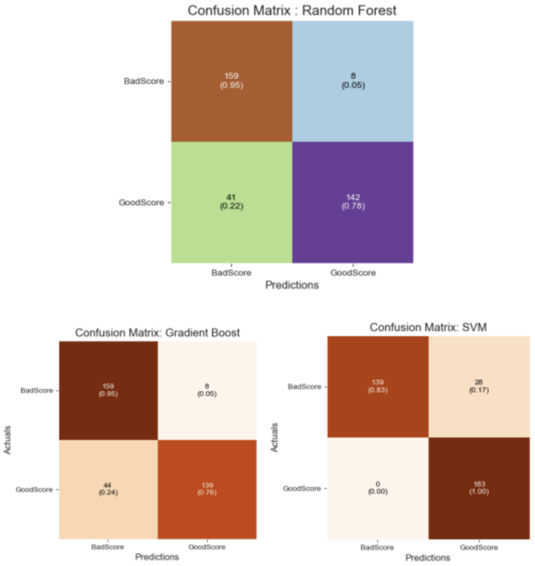

Fig. 4. Confusion Matrix of over-sampled data

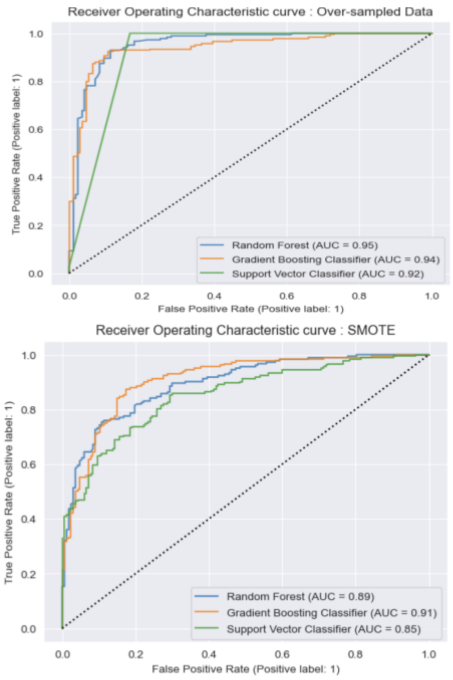

Fig. 5. ROC-AUC Curve

VI. DISCUSSION AND CONCLUSION

The high accuracy rate of 10-fold cross validation offers reassurance that the models have mostly correctly identified the patterns in the data and are low in bias and variance. Overall, the score of oversampled data is higher than SMOTENC, indicating that they outperformed SMOTENC in terms of outcomes. This may be because of the SMOTENC technique's tendency to produce irrational samples that aren't always representative of the minority class.

The major goal of bank credit scoring is to prevent loan defaults by accurately identifying applicants who face a high risk of defaulting on their loans; as a result, false negatives are significantly more destructive in this situation than false positives. Too many false positives would indicate that the bank is losing too many profitable clients and too many false negatives would mean that the bank is suffering losses as a result of granting risky loans. Hence, in this model we need to prioritize recall (also known as sensitivity), over precision.

Support Vector machine on oversampled data have the highest accuracy rate of 92% and has 100% false positive recall (Fig.4), which means it didn’t misclassify any good customers as bad. However, the high (17%) false-negative recall rate makes this model less efficient. Hence, although high in accuracy, SVM model performed the least among the three and the same can be observed in the first image of the ROC-AUC Curve in Fig.5.

Random Forest model and Gradient boost classifier on oversampled data performed the best because it balances precision and recall effectively (as shown in Figure 4) while keeping acceptable accuracy.

VII. APPENDIX

The following Google Drive link includes the original data as well as the Python code used to pre-process the data and apply the categorization techniques : https://livecoventryac-my.sharepoint.com/:f:/g/personal/nairm2_uni_coventry_ac_uk/EtnSOgjvyShBlHVW5OKfiT0BYh4948awWuEMFnU64kIiqQ?e=pWdZVE

REFERENCES

[1] Khan, Mausikar. (2018). *Global Financial Crisis of 2008 and its Impacts*. https://www.researchgate.net/publication/349502501_Global_Financial_Crisis_of_2008_and_its_Impacts

[2] Durand, D. (1941). *Risk elements in consumer instalment financing*. National Bureau of Economic Research.

[3] Çallı, B. A., & Coşkun, E. (2021). *A Longitudinal Systematic Review of Credit Risk Assessment and Credit Default Predictors*. SAGE Open, 11(4). https://doi.org/10.1177/21582440211061333

[4] Capon, N. (1982). Credit Scoring Systems: A Critical Analysis. *Journal of Marketing, 46*(2), 82–91. https://doi.org/10.1177/002224298204600209

[5] Bumacov V., Ashta A., Singh P. (2017). Credit scoring: A historic recurrence in microfinance. *Strategic Change, 26*(6), 543–554. https://doi.org/10.1002/jsc.2165

[6] Onay C., Öztürk E. (2018). A review of credit scoring research in the age of Big Data. *Journal of Financial Regulation and Compliance, 26*(3), 382–405. https://doi.org/10.1108/jfrc-06-2017-0054

[7] ICO. (2017). *Big data, artificial intelligence, machine learning and data protection.* https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

[8] Min, J. H., & Lee, Y. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications, 28*(4), 603-614. doi:10.1016/j.eswa.2004.12.008

[9] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[10] Meltzer, R. (2021, July 15). *What is Random Forest?*. CareerFoundry. Retrieved February 5, 2023, from https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/#:~:text=Random%20forest%20is%20used%20on,%2C%20patient%20history%2C%20and%20safety.%2012)%20https://blog.paperspace.com/gradient-boosting-for-classification/

[11] Kurama, V. (2020). *Gradient Boosting In Classification: Not a Black Box Anymore*. PaperspaceBlog. Retrieved February 7, 2023, from https://blog.paperspace.com/gradient-boosting-for-classification/

[12] Tokuc, A.A. (2022). *Why Feature Scaling in SVM?.* Baeldung. Retrieved February 7, 2023, from https://www.baeldung.com/cs/svm-feature-scaling

[13] Aguilar, F. (2019, October 9). *SMOTE-NC in ML Categorization Models for Imbalanced Datasets*. Medium. Retrieved February 1, 2023, from https://medium.com/analytics-vidhya/SMOTE-nc-in-ml-categorization-models-fo-imbalanced-datasets-8adbdcf08c25

[14] Yangyudongnanxin, G. *Financial Credit Risk Control Strategy Based on Weighted Random Forest Algorithm.* Scientific Programming. https://doi.org/10.1155/2021/6276155

[15] Tian, Zhenya & Xiao, Jialiang & Feng, Haonan & Wei, Yutian. (2020). *Credit Risk Assessment based on Gradient Boosting Decision Tree.* Procedia Computer Science. 174. 150-160. 10.1016/j.procs.2020.06.070.

[16] Gupta, L. (2020, November 21). *Comparison of Hyperparameter Tuning algorithms: Grid search, Random search, Bayesian optimization*. Medium. Retrieved February 1, 2023, from https://medium.com/analytics-vidhya/comparison-of-hyperparameter-tuning-algorithms-grid-search-random-search-bayesian-optimization-5326aaef1bd1