Assignment Title: Reflective Report

Module: BAA1027

Lecturer: Michael Farayola

Student Name: Liam Cummins

Student Number: 22111611

Word Count: 3,070 words

Table of Contents:

Reflective Report: **Predicting Bank Loan Approvals with the support of Machine Learning Models**

1. Introduction

I specifically chose this issue because accurate predictions in bank loan approvals have become an essential aspect of modern finance, with many banks and other financial institutions striving to reduce likelihood of lending to risky applicants, while making faster, more well-thought-out decisions in money lending endeavours. This project aimed to explore how machine learning models can assist in predicting whether a loan applicant is reliable in their abilities to repay a loan, based on data collected in the machine learning models.

I chose this topic due to its real-world applicability and that it combines the topics of finance and data science, both of which are areas I am passionate about, and along with having previous experience in the finance sector. The main goal was to evaluate different machine learning models and identify which performed best for this classification task, as banks depend on reliable loan applicants to pay back loans as their main source of revenue.

Before committing to this assignment, I expected classification models such as Logistic Regression to perform well due to not being overly difficult to understand and that it could be a good starting point, before moving onto, more complex models being Random Forests or Support Vector Machines (SVM) to see how well these models perform compared to the simpler model of linear regression.

2. Related Work

The use of machine learning in financial matters has significantly increased and become more common throughout the decade, especially concerning credit risk and loan approvals. Many financial institutions prepare datasets of their clients for cleaning and sort them into a choice of classification models to help support lending decisions.

A typical supervised learning model used in loan applications are logistic regression models as these models are known for being robust and simpler to understand, However, this model has started to see less use, mainly due to it not being able to capture non-linear values and with banks requiring more complex models for a more effective analysis. (Nureni & Adekola, 2022). Recently most financial institutions have made the move to integrate more complex such as decision trees and ensemble methods who can potentially outperform classic models due to their superior performance and ability to model complex datasets. (YILDIZ, 2023)

Ensemble methods such as random forest and gradient boosting were shown to provide the most consistent highest levels of accuracy and are versatile with other methods such as decision trees, making them essential for multiple datasets.(Arun,Ishan, &Sanmeet,., 2016) This is backed by many research papers who results favour ensemble methods algorithms, while traditional classifiers such as linear regression underperformed and results were skewed in this case of class distributions, an essential part of financial applications. (Uddin, et al, 2023)

With class imbalance being a significant limitation for this dataset, I chose to address this with the use of SMOTE (Synthetic Minority Oversampling Technique), to better improve generalization throughout this model, along with mitigate bias surrounding the majority class(approved) by creating synthetic data to be supplemented for the minority class(rejected). (Blagus &Lusa, 2013). The SMOTE technique can potentially help in evaluating each machine learning model for fairer results.

This review helped shape my research, for example, realizing that many high-performing models lack transparency influenced my decision to compare them against simpler models. It also made me conscious of potential bias in feature selection and data representation, encouraging a more ethical lens throughout the project.

3. Methodology

The dataset for this project was sourced from Kaggle and included various demographic and financial attributes, such as applicant income, education, credit history (cibil score), loan amount and loan status. The dependent variable was binary, as it dictated whether a loan was to be approved or rejected. Upon inspection, the dataset exhibited several challenges: missing values, categorical variables, duplicate rows, and class imbalance.

Data Cleaning& Preprocessing

To handle these issues, I carried out the command to remove duplicates. Then handle missing values by starting the imputation process, where missing numerical values were replaced using mean and median, while missing categorical data was replaced using the mode function. The column names were stripped of trailing whitespaces and renamed for better clarity.

This dataset is heavily continuous values, as such I employed the binning method of one hot encoding to transform categorical variables into discrete numerical data. The ID column was of no valuable use to the model, so this column was dropped, along with this target and features were defined. Last step a train test split was applied to ensure no risk of overfitting, thus making it most highly useable for machine learning models.

Addressing Class Imbalance

Data imbalance was a noticeable problem in my classification models, as the dataset heavily favored approval of loans over rejections. To have a better understanding of the class distribution I created a pie chart, to better visualize this imbalance.

Given how great of an imbalance there is, I implemented SMOTE to generate synthetic data to be supplemented for the minority class, being rejected loan applicants (Blagus& Lusa, 2013). By introducing SMOTE to the dataset, it significantly improved f1 score and recall, especially in underperforming models like Support Vector Machine (SVM), where class imbalance dictated its results.

Model Selection and Reasoning

For this dataset I chose four supervised learning classification models to evaluate:

- **Logistic Regression**: It is well known being highly readable compared to other models, while also having a good balance of performance. (PRIYA & KUMARI, 2024; Nureni & Adekola, 2022) I thought that it would be a good starting point to help with creating more complex models as I progress.
- **Decision Tree**: It is a well-structured tree-based model, capable of capturing non-linear and complex relationships and can also be used to further enhance other models like ensemble methods. (Alagic, et al, 2024)
- **Random Forest**: built on decision trees, it eliminates correlation through its random sampling, it is also versatile enough to work with large datasets, especially if these datasets have missing data or outliers, being effective at generalization (Alagic, et al, 2024)
- **Support Vector Machine (SVM)**: This algorithm works very well with high dimensional data, but with the drawback of being extremely sensitive when the dataset has a class imbalance, meaning that SMOTE implementation could be vital for this model (Li, Shiue, & Huang, 2006).

Evaluation Metrics

Each model was evaluated with the implementation of a confusion matrix to see actual and predicted results together, for a better understanding of each models performance and possible classification problems.(Deng, et al., 2016).Evaluation metrics included accuracy, precision, recall, F1-score, along with an ROC (Receiver Operating Characteristic) curve, to better visualize how well each model did compared to each other. (Fawcett, 2006). Luckily, no model was not in danger of being in AUC (Area under the curve), indicating no severe classification bias for any models (Fawcett, 2006).

Reflection

Reflecting on the development phase, data preprocessing was more time-consuming than anticipated, with the initial phase to create a logistic regression model requiring more time to complete correctly without any errors, especially concerning class imbalance. However, the hard work paid off to build a foundation, along with having a better understand for implementing other models going forward.

I gained practical insights into the importance of thoughtful feature engineering and the potential dangers of overlooking data bias, and the challenges that arise from the trade-off between interpretability and high performance. While random forests offer the best results, the model lacks simple readability, which is a crucial factor for financial institutions in choosing models for depicting financial concerns.

4.Results and Analysis

Each model was evaluated using accuracy, F1-score, precision, and recall ensuring a balanced and fair assessment, which is particularly important due to the moderate class imbalance (62% approved, 38% rejected loans). The models were tested before and after applying SMOTE for clearer results and insights into the intervention of this technique.

4.1 Logistic Regression

Before applying SMOTE, Logistic Regression achieved an accuracy of 79.9% and a weighted F1-score of 0.79. However, the recall for rejected loans (class 0) was only 0.61, indicating the model mistook rejections as approvals. After SMOTE, the model slightly declined in overall accuracy to 78.5% but improved class 0 recall to 0.68, suggesting a fairer representation of the minority class.

Despite being the least complex model, logistic regression’s simpler interpretability helps it stand out to financial services who desire transparency and effective performance in the model. (Nureni & Adekola, 2022)

4.2 Decision Tree

The Decision Tree classifier performed effectively under both conditions. Before SMOTE, it achieved 97.3% accuracy and a weighted F1-score of 0.97, showing it as a strong model for class-based precision and recall. After SMOTE, performance marginally improved to 97.8% accuracy and a 0.98 F1-score, with only 19 misclassifications out of 854 predictions.

These results highly favour Decision Trees as a suitable model for loan applications, as the model has an effective structure, is robust enough to prevent outliers, curbing problems with class imbalance problematic throughout this dataset. (Nureni & Adekola, 2022) (Alagic et al., 2024).

4.3 Support Vector Machine (SVM)

SVM has performed as the worst model throughout this dataset, achieving only 62.8% accuracy and a weighted F1-score of 0.48. The model completely failed to recognise any rejected loans, as class 0 was 0% values for recall and f1-score, meaning it ineffectively predicted all observations as approved. However, after SMOTE, recall for class 0 improved to 0.45, while F1-score improved to 0.54, though accuracy fell to 53.6%, sacrificing it for an improved performance.

While SMOTE helped improved SVM this is still not a good outcome and presents SVM as the weakest model for loan application predictions. SVM should not be considered for bank loan predictions in the future as the model is not suited for large datasets, it is prone to class imbalance and can put computational strain on your computer compared to the more effective models outlined. (Nureni & Adekola, 2022).

4.4 Random Forest

Random Forest consistently outperformed all other models. Before SMOTE implementation, it achieved 97.9% accuracy and a 0.98 F1-score. After SMOTE, results showed no significant changes, with 98.1% accuracy and an unchanged F1-score, along with high precision and recall across both classes.

Random forest is shown to be the best performing of the models implemented, likely due to it being well suited for classification tasks as the model is effectively at dealing with large imbalanced datasets. (Nureni & Adekola, 2022)

4.5 Visual & Ethical Reflections

Visual tools such as ROC curves and confusion matrices clarified model performance and trade-offs. Random Forest showed the best results in the ROC curve, with a near-perfect AUC, while SVM’s ROC curve performed least effectively as it was incredibly close to being below the AUG curve, indicating poor class separation

Effective performance is crucial in bank loan predictions as false positives (high risk applicant) approval could result in economic loss. While false negatives (low risk applicant) mean individuals are denied loans unjustifiably and the bank loses high potential lenders to drive up profit. This highlights that it is important to use an effective model for loan purposes but not being dependent on automated decisions and instead allow your critical thinking skills to assess fairness and equality in loan approvals.

4.6 Impact of SMOTE

The application of SMOTE as part of this dataset has had a varying impact to each model:

- Logistic Regression saw a recall increase for class 0 (from 0.61 to 0.68) and a small dip in overall accuracy, indicating better balance in prediction but limited impact on overall performance.
- Decision Tree and Random Forest showed only minor changes, as tree-based models are naturally more robust for larger classification datasets problems contributing to their split-based learning approach. (Nureni& Adekola, 2022).

- SVM saw the most drastic change, moving from total failure on class 0 (0% recall) to 45% recall after SMOTE. However, its overall accuracy and F1-score remained the weakest, showing that oversampling helped basic classification but not decision boundaries.

These observations reinforce that while SMOTE is a useful tool for balancing training data, it does not guarantee equal improvements across all models. The choice of classification model significantly influences the role it plays in loan applications, requiring critical aspects such as accuracy, fairness, interpretability, and accuracy.

5. Discussion

5.1 Interpretation of Results

Amongst the models evaluated, Random Forest proved the to be the most consistently effective for loan application predictions, the is likely due to the model being remarkably effective with working on high dimensional data that it can generalize for classification tasks. (Petropoulos, et al, 2020) (Rodriguez-Galiano,et al, 2015). This supports prior research that suggests ensemble methods are the most effective models in the implementation of loan application predictions and other areas concerning finance. (Chopra & Bhilare, 2018)

The Decision Tree also performed exceptionally, with a small risk of overfitting. Even so its results align with those from previous academic papers and sees truly little challenge of equal transparency and performance.

In contrast, Logistic Regression, accuracy was not its best aspect, however its simpler approach allowed for more easier interpretability and transparency that is essential in financial institutions.

Support Vector Machine (SVM) performance was the weakest, especially before SMOTE was applied, as it failed to predict the entirety of the rejected loans. Even after SMOTE, class 0 recall only rose to 0.45. This reinforces that SVM is not the best model to handle large datasets (Nureni& Adekola, 2022).

5.2 Limitations

This project had several limitations, one of which would be the size of the dataset. The overall size is minor compared to real world financial datasets, which in-turn limits the potential of the dataset to generalize, leading to the necessity of SMOTE to create new synthetic data to help capture a more effective loan application rate.

Secondly, the models were influenced by the default hyperparameters, without performing hyperparameter tunning or cross-validation, by relying on default setting we limit the potential of our models, especially SVM, as they are overly sensitive to hyperparameters. (Bergestra& Bengio,2012). Going forward, projects would benefit from the intervention of grid search and random search to improve accuracy and generalization throughout the model without unfair influencing factors. (Bergestra& Bengio,2012; Belete & Huchaiah, 2022) Future

Furthermore, the models are limited to one dataset, when additional dataset such as behavioural data (e.g., spending habits), which are commonly used in real-world credit scoring, could have further enhanced the dataset. By incorporating more features from an additional dataset, we can increase the realism of the models, while also having more meaningful results.

5.3 Ethical and Practical Considerations

Financial Institutions put more stock into models with fairness and interpretability, rather than effective performance. In this case, Random Forest had the best performance rate of the models used, but it is challenged due to its deep complexity, making too hard to read for those not fluent in machine learning.

In contrast, Logistics Regression, performed well, despite not being the most optimal of the models. However, many favour this model due to it being less complex, leading to better transparency, making it be seen as the best model for decision-making. (Nureni& Adekola, 2022).

Another concern is bias in this dataset, features such as, marital status and education can influence the decision to approve or reject loan applications. For example, a married, educated person would be more likely approved as they are seen as a reliable and stable lender. This positive discrimination hurts both sides as these factors should not influence decision making on the side of the bank and it is unintendedly saying those who are single and uneducated are unfit choices for loans.

Misclassification can also come with a social cost. While false positives and negatives impacts a bank's financial stability, the choice of denying this essential loan can have a massive impact on a person's financial well-being, forgetting the human need in these loans. In this case machine learning models should not replace human decision making and instead use it to help enhance decision making.

5.4 Future Directions of Machine Learning

Throughout this project, I have learned about the significant impact issues such as class imbalance can have on the performance of our models and the necessity of machine learning tools in enhancing fairness for each model independently.

In future work, I would:

- Implement hyperparameter tunning such as grid search and random search to help elevate the results of models influence immensely by default set hyperparameter, such as SVM (Bergestra& Bengio,2012; Belete & Huchaiah, 2022).
- Incorporate explainability tools such as SHAP (SHapley Addictive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), these tools can help enhance the interpretability of overly complex models such as random forest, so that financial institutions are not influenced by the trade-off between interpretability over performance. (Salih, et al, 2025.)
- Experiment with more models, specifically ensemble methods as random forests had consistently high results. Integration of ADA Boost could be beneficial as the model is known for making use of weak classifiers and making a more efficient dataset, despite it being imbalanced. (Sayed, et al, 2024)

This project has taught me about the reasoning behind selecting certain models for datasets. With this in mind I will not stray away from using a dataset due to performance or interpretation but instead consider what I could do to address these shortcomings to make a more efficient model, while also being open to lesser-known models who could also have a significant impact on future projects.

6. Conclusion

The purpose of this project was to evaluate the effectiveness of machine learning models in the use of bank loan predictions to determine approval and rejections. By comparing Logistic Regression, Decision Tree, Support Vector Machine, and Random Forest models, both before and after applying SMOTE for misclassification issues, the project aimed to explore not just predictive accuracy, but also fairness, interpretability, and practical implication of each model.

Random Forest was consistently the best in its performance regardless of SMOTE. Decision Trees also performed well, while Logistic Regression performed optimally but is more useful in interpretability. However, SVM had the poorest results of all four models, with class imbalance deeply affecting its performance, reinforcing the need for hyperparameter tuning and model selection in real-world applications.

Throughout this project, I gained practical experience in model evaluation, class balancing, and ethical analysis. I also developed a deeper appreciation for the complexity of integrating machine learning in business areas such as financial loans, where transparency, trust, and accountability are just as important as the likelihood of the loan being repaid.

Bibliography

Kaggle Dataset Link: https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset

madandahal. (n.d.) Auto MPG – Linear Regression [Jupyter notebook]. GitHub. Available at: https://github.com/madandahal/Auto-Mpg--Linear-Regression/blob/master/.ipynb_checkpoints/Auto%20MPG_Senior-checkpoint.ipynb (Accessed: 3 April 2025).

Alagic, A., Zivic, N., Kadusic, E., Hamzic, D., Hadzajlic, N., Dizdarevic, M. and Selmanovic, E., 2024. Machine learning for an enhanced credit risk analysis: A comparative study of loan approval prediction models integrating mental health data. Machine Learning and Knowledge Extraction, 6(1), pp.53-77.

Arun, K., Ishan, G. and Sanmeet, K., 2016. Loan approval prediction based on machine learning approach. IOSR J. Comput. Eng, 18(3), pp.18-21.

Belete, D.M. and Huchaiah, M.D., 2022. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. International Journal of Computers and Applications, 44(9), pp.875-886.

Bergstra, J. and Bengio, Y., 2012. Random search for hyper-parameter optimization. The Journal of Machine Learning Research, 13(1), pp.281-305.

Blagus, R. and Lusa, L., 2013. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics, 14, pp.1-16

Chopra, A. and Bhilare, P., 2018. Application of Ensemble Models in Credit Scoring Models. Business Perspectives and Research, 6(2), pp.129–141. https://doi.org/10.1177/2278533718765531.

Deng, X., Liu, Q., Deng, Y. and Mahadevan, S., 2016. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Information Sciences, 340, pp.250-261.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27(8), pp.861-874.

Li, S.T., Shiue, W. and Huang, M.H., 2006. The evaluation of consumer loans using support vector machines. Expert Systems with Applications, 30(4), pp.772-782.

Nureni, A.A. and Adekola, O.E., 2022. Loan approval prediction based on machine learning approach. FUDMA Journal of Sciences, 6(3), pp.41-50.

Petropoulos, A., Siakoulis, V., Stavroulakis, E. and Vlachogiannakis, N.E., 2020. Predicting bank insolvencies using machine learning techniques. International Journal of Forecasting, 36(3), pp.1092-1113.

PRIYA, T.K.S. and KUMARI, N.L., 2024. LOAN APPROVAL PREDICTION USING MACHINE LEARNING. Journal of Nonlinear Analysis and Optimization, 15(2).

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M. and Chica-Rivas, M.J.O.G.R., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71, pp.804-818.

Salih, A.M., Raisi-Estabragh, Z., Galazzo, I.B., Radeva, P., Petersen, S.E., Lekadir, K. and Menegaz, G., 2025. A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7(1), p.2400304.

Sayed, E.H., Alabrah, A., Rahouma, K.H., Zohaib, M. and Badry, R.M., 2024. Machine Learning and Deep Learning for Loan Prediction in Banking: Exploring Ensemble Methods and Data Balancing. IEEE Access.

Uddin, N., Ahamed, M.K.U., Uddin, M.A., Islam, M.M., Talukder, M.A. and Aryal, S., 2023. An ensemble machine learning based bank loan approval predictions system with a smart application. International Journal of Cognitive Computing in Engineering, 4, pp.327-339.

YILDIZ, A., 2023. Determining the factors for individual credit approval by applying logistic regression and hierarchical logistic regression. International Journal of Management Studies and Social Science Research, 5(6), pp.58-67.