POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Advancing Loan Default Prediction with Interpretable TabNet Models

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Arnaldo Mollo

Advisor: Prof. Daniele Marazzina

Academic year: 2022-2023

1. Introduction

This master thesis offers an in-depth study of loan default prediction using sophisticated machine learning models, with special emphasis on the TabNet model. This work addresses a critical issue in the field of machine learning - the trade-off between the high predictive power of black-box models and the transparency and interpretability offered by white-box models. It is in this context that TabNet is analyzed, as it successfully bridges this gap, marrying high accuracy with interpretable outputs.

We build on the research foundations laid by Cascarino, Moscatelli, Parlapiano, and others [2], focusing our investigation on the performance and interpretability of the TabNet model. Our study undertakes a comparative analysis of several machine learning models including LightGBM, XGBoost, Logit, and Random Forest. However, the cornerstone of our work is the in-depth exploration and critical analysis of the TabNet model's interpretability.

In the course of our work, we critique model-agnostic interpretability techniques such as SHAP values [4], highlighting their limitations and computational cost. We argue for the advantage of TabNet's intrinsic interpretability, demonstrated through a detailed investigation of TabNet's masks - a feature importance mechanism that offers nuanced insights at each decision step of the model.

Our analysis uses a comprehensive suite of performance metrics to show that the TabNet model strikes an optimal balance between prediction accuracy and interpretability. This in-depth investigation of the TabNet model's masks provides unique insights into the complex interplay and influence of features at various stages of the decision-making process, a level of understanding that is invaluable in intricate domains such as credit risk management.

The thesis makes a significant contribution to the existing body of research in loan default prediction by providing practical insights that will guide the financial industry towards better-informed credit risk assessment and management decisions. It is structured into six chapters that cover a wide spectrum of topics, including existing models and interpretability techniques, a thorough dissection of the TabNet model and its architecture, a presentation of the dataset and preprocessing methodologies, a detailed exposition of the training of different models and their results, and thorough interpretability analyses for both TabNet and LightGBM. The thesis concludes by summarizing the key findings and suggesting potential avenues for future research.

2. The TabNet Model

The TabNet model, proposed by Arik et al. [1], is a ground-breaking approach to tabular data analysis, bridging the gap between traditional machine learning methods and contemporary deep learning techniques. It is inspired by transformers [5] and incorporates a unique architecture combining learnable sparse feature selection, attention mechanisms, and end-to-end training to tackle structured data challenges such as missing values, categorical variables, and complex feature interactions.

The critical elements of the TabNet architecture include:

1. **Feature Embeddings:** TabNet transforms raw tabular data through an embedding layer, creating a continuous representation that allows for the efficient processing of both numerical and categorical variables, eliminating the need for separate preprocessing.
2. **Sequential Attention and Feature Selection:** At each decision step, TabNet employs an attention mechanism to focus on the most salient features. This instance-wise feature selection boosts the model's interpretability and learning efficiency.
3. **ReLU Activation and Element-wise Addition:** Non-linearity is introduced via a ReLU activation function, and information from each decision step is aggregated through element-wise addition.
4. **Fully-Connected Layer (FC):** The model incorporates an FC layer to model complex relationships and capture non-linear patterns in the data.
5. **Output Layer:** The final representation is processed through an output layer to generate the model's predictions, based on the specific task.
6. **Autoencoder and Unsupervised Training:** The TabNet architecture supports unsupervised training via an autoencoder, using unlabeled data to learn meaningful representations before fine-tuning with supervised training on labeled data. This two-phase training approach improves performance and generalization, particularly when labeled data is scarce or noisy.

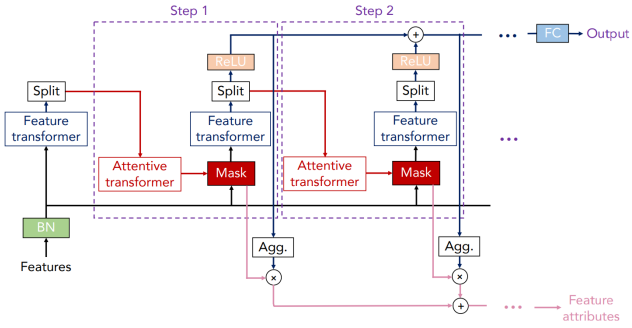

Figure 1: The Transformer Architecture

In conclusion, TabNet offers a powerful, scalable, and interpretable solution for processing and analyzing structured data, demonstrating competitive performance across various benchmark datasets. It represents the perfect marriage of traditional and deep learning techniques in a model designed specifically for tabular data analysis.

3. The Lending Club Dataset

The data used is derived from the Lending Club dataset, a comprehensive source of loan data spanning from 2007 to 2018. The dataset contains over 2.2 million samples, each representing an individual loan and characterized by 151 features, both numerical and categorical.

This resource has been widely used in both academic and industrial contexts for credit risk assessment and loan default prediction, providing a versatile platform for exploration in these areas. However, it is not without challenges. There is significant presence of missing data (31.78%), requiring strategic handling during preprocessing. Additionally, features with high levels of missing data (over 90%) have been removed to prevent introduction of noise and excessive dimensionality.

Among the numerical features, 'loan_amnt' is of particular interest as it represents the loan amount for each sample. It is directly tied to a borrower's financial obligation and their capacity to repay. The loan amount varies from as low as $500 to as high as $40,000, with an average value of around $15,046.93. The standard deviation of approximately $9,190.25 reflects the substantial dispersion in loan amounts across borrowers, see figure 2.

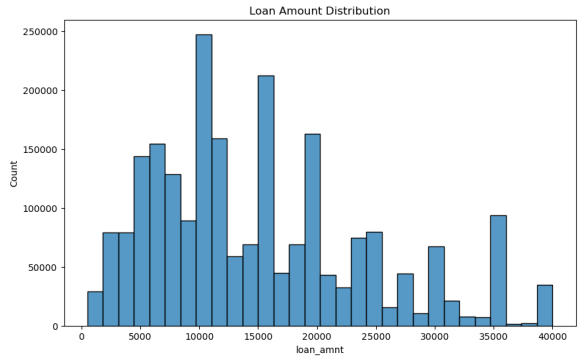

Figure 2: Loan Amount Distribution

Diving deeper, the loan amount is further dissected by the borrower's grade, categorized from A to G. The borrower's grade is an assessment of the borrower's creditworthiness, with A being the highest grade and G the lowest. Intriguingly, the mean loan amount tends to increase as the grade worsens. This observation underlines the riskier nature of lower-grade borrowers, who typically face higher interest rates due to increased credit risk, figure 3.

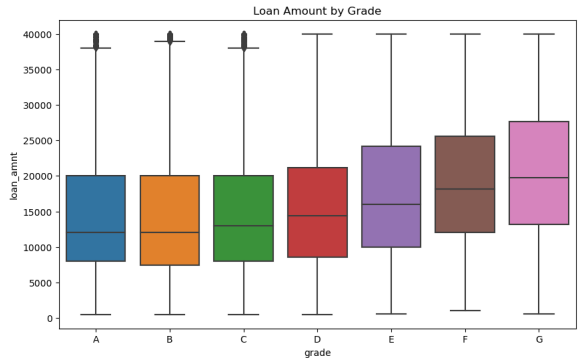

Figure 3: Loan Amount by Grade

In terms of feature preprocessing, the initial focus was on dealing with missing values. For categorical features, missing values were filled with the string 'missing', while for numerical features, they were replaced with -1. This preserves the original data distribution and prevents potential issues with machine learning algorithms, which are typically not designed to handle missing data.

The dataset was further processed to avoid potential data leakage, an issue that arises when information not available at the time of prediction is inadvertently included in the training data. Certain features that could cause such leakage were identified and removed.

Next, high-dimensional categorical features were also considered, which can pose challenges due to their potential to increase model complexity, contribute to overfitting, and decrease computational efficiency. Noisy or misleading high dimensional categorical features were removed from the dataset.

Finally, the encoding of labels was addressed, with an aim to simplify loan statuses for the machine learning models to process more efficiently. The statuses were simplified to a binary outcome: 0 for 'Fully Paid' loans, and 1 for 'Charged Off' and 'Default' loans.

As a whole, this dataset, while challenging, provides a solid foundation for the construction of a predictive model for loan defaults. Proper preprocessing ensures robustness, efficiency, and accuracy of the models developed. The preprocessing also maintains the real-world applicability and interpretability of the data.

4. Unsupervised and Supervised Training

In this analysis, five machine learning models - LightGBM, XGBoost, TabNet, Random Forest, and Logistic Regression - were trained for loan default prediction. Performance was compared using several metrics, such as Validation Accuracy, Test Accuracy, Test Precision, Test Recall, Test F1-score, and Test AUC-ROC score.

TabNet demonstrated superior results in Validation Accuracy (0.675) and Test Accuracy (0.673), meaning it had the highest proportion of correct predictions in both validation and test datasets. It also scored highest in Test Precision (0.334), indicating its predictions of loan default are slightly more reliable than other models.

However, in terms of Test Recall, LightGBM led with a score of 0.693, illustrating better ability in identifying all actual positive instances. LightGBM and XGBoost also tied for the highest Test F1-score (0.449), suggesting a more effective balance between precision and recall. Furthermore, LightGBM achieved the highest Test AUC-ROC score (0.738), a critical measure of the model's ability to distinguish between positive and negative classes.

While TabNet did not top all metrics, it performed competitively across the board while also offering superior interpretability, making it a compelling choice for tasks that require both

high performance and explainability, such as loan default prediction. The performances of Random Forest and Logistic Regression were somewhat lower, especially in the Test AUC-ROC score, indicating their lesser ability to distinguish between loan default and non-default cases.

Overall, this study highlights the potential of TabNet for tasks that demand both performance and model interpretability, providing valuable insights for those applying machine learning models for loan default prediction.

Metric	LGBM	XGB	TNet	RF	Logit
Val. Acc.	0.660	0.660	**0.675**	0.654	0.663
Test Acc.	0.660	0.661	**0.673**	0.653	0.663
Test Prec.	0.332	0.333	**0.334**	0.323	0.329
Test Rec.	**0.693**	0.691	0.641	0.668	0.656
Test F1	**0.449**	**0.449**	0.439	0.435	0.438
Test AUC	**0.738**	0.737	0.723	0.659	0.660

Table 1: Metrics for LGBM, XGB, TabNet, RF, and Logit

5. Interpreting Predictive Model Decisions

In this study, we dissected the mechanisms behind the decisions of a predictive model for loan repayment. We evaluated two distinct methodologies for this purpose: SHAP values and TabNet Masks. Their application provided insights into the feature importance and influence on the model's decision-making process.

5.1. Global Interpretability

The global interpretability analysis provides an all-encompassing view of the feature importance in the prediction process of the model. While usually this is achieved by measuring the extent to which variations in the input features' values impact the model's output, TabNet has a unique advantage when it comes to interpretability. Its attention mechanism allows us to track which features the model is focusing on at each decision step, providing transparency into the model's internal workings. The mask values, ranging from 0 to 1, indicate the degree of attention given to each feature at each decision step. In this context, we investigate the global importance of features in the TabNet model, trained on the Lending Club dataset. The aggregate feature importance is shown in figure 4

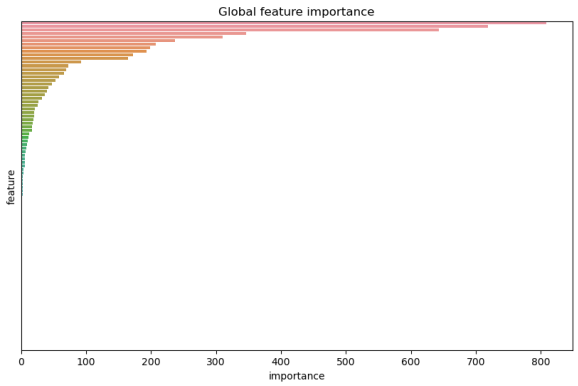

Figure 4: Features Global Importance

5.2. Case Studies and Important Features

Our analysis involved three individual borrowers, each with unique financial profiles and loan repayment predictions. The decision-making model considered a multitude of features, which varied in their influence over the predicted outcome.

For the first borrower, the model anticipated a loan default. In figure 5 we can see the tabnet masks associated with him.

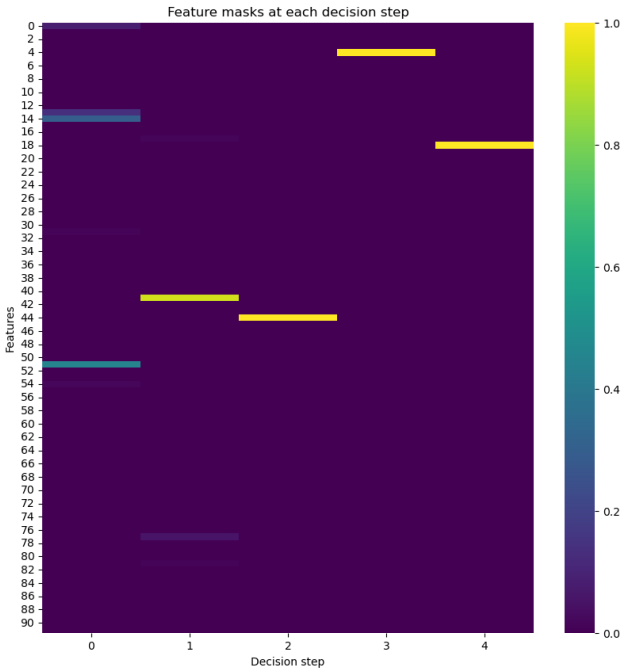

Figure 5: Borrower 1 masks at Each Decision Step

Key contributors to this prediction were their high-risk grade (G3) and short employment his-

tory. The second borrower, with the highest grade (A1) and a loan purpose of credit card consolidation, was predicted to fully repay the loan. This was likely due to the lower risk associated with their grade and responsible credit behavior. Finally, the third borrower was also expected to fully repay their loan. Despite their middle-risk grade (C5), other features such as a solid FICO score and responsible financial behavior suggested a lower risk.

5.3. SHAP Values Methodology

SHAP values, Lundberg 2017[3], based on principles from cooperative game theory, ensures a balanced allocation of feature importance by considering all possible feature combinations. This comprehensive approach is applicable across a variety of models, providing consistent and fair interpretations.
Mathematically, the Shapley value for a feature i is defined as:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}(f(S \cup \{i\}) - f(S)), \quad (1)$$

Where, f represents the prediction function, N denotes the set of all features, and S is any subset of N excluding feature i. The cardinalities of S and N are represented by $|S|$ and $|N|$ respectively. The Shapley value $\phi_i(f)$ calculates the average marginal contribution of feature i over all possible subsets of features, summing up the differences in the prediction function's output when feature i is included and excluded from each subset.
The exact computation of SHAP values is a non-trivial task. The SHAP explanation model simplified input mapping is then given as:

$$f(h_x(z^0)) = E[f(z)|z_S], \quad (2)$$

Where, f is the function that the SHAP values aim to explain, h_x is a simplified input mapping function, z^0 represents the simplified version of the instance, $E[f(z)|z_S]$ represents the expected value of $f(z)$ given the features in the set S, z_S represents the values of features in the set S and S is a subset of all features.
However, there are a few limitations. SHAP values assume the independence of features, which might not hold true in real-world scenarios, where correlations between features often exist. In addition, it can be computationally intensive, particularly with large feature spaces, which can pose challenges in real-time systems where computational speed is crucial. Moreover, by focusing on the average contribution of a feature, it doesn't explicitly account for feature interactions, potentially obscuring complex decision-making processes.

5.4. TabNet Masks Methodology

TabNet Masks, on the other hand, assigns feature importance at each decision step within the TabNet model. This approach provides a granular understanding of how features interact and influence predictions at different stages of the process.
The aggregate decision contribution at the i^{th} decision step for the b^{th} sample, denoted as $\eta_b[i]$, is computed by applying the formula:

$$\eta_b[i] = \sum_{c=1}^{N_d} ReLU(d_{b,c}[i]), \quad (3)$$

In this formula, the Rectified Linear Unit (ReLU) activation function is used, defined as $ReLU(x) = max(0, x)$. This function is applied on the decision step output $d_{b,c}[i]$, contributing to the aggregate value $\eta_b[i]$.
The aggregate feature importance mask is proposed in the paper as a way to weigh the relative importance of each feature in the decision-making process. The formula is given as:

$$M_{agg-b,j} = \frac{\sum_{i=1}^{N_{steps}} \eta_b[i] M_{b,j}[i]}{\sum_{j=1}^{D} \sum_{i=1}^{N_{steps}} \eta_b[i] M_{b,j}[i]}, \quad (4)$$

In this formula, $M_{agg-b,j}$ symbolizes the aggregate feature importance mask for the j^{th} feature of the b^{th} sample. The term $\eta_b[i]$ is the aggregate decision contribution at the i^{th} decision step for the b^{th} sample, as we defined earlier. The symbol $M_{b,j}[i]$ represents the feature selection mask for the j^{th} feature of the b^{th} sample at the i^{th} decision step. Finally, D signifies the total count of features, while N_{steps} corresponds to the overall number of decision steps.
Finally, to obtain a global feature importance measure, we sum the aggregate feature importance masks across all samples and normalize by 1000. This gives a measure that reflects the average importance of each feature across all samples in the dataset, thereby providing a global view

of feature importance. The formula for this is given as:

$$G_j = \frac{1}{1000} \sum_{b=1}^{N_{samples}} M_{agg-b,j}, \qquad (5)$$

where G_j is the importance of feature j globally. $M_{agg-b,j}$ represents how much feature j matters in sample b. $N_{samples}$ is the total count of samples.

Unlike SHAP values, TabNet Masks can reveal the complex interactions and sequential importance of features, contributing to a richer understanding of the decision-making process. This capability is especially useful in intricate domains such as credit risk, where understanding the sequence and interaction of feature importance can provide crucial insights.

6. Conclusion

This thesis presents an in-depth study into the application of TabNet, a highly interpretable machine learning model, in predicting loan defaults. It demonstrates that TabNet not only competes well against other models in terms of Validation Accuracy, Test Accuracy, and Test Precision but also provides a balanced compromise between performance and interpretability. Of particular note is its superior precision, which reduces the risk of costly false positives prevalent in financial applications.

The study delves into a comprehensive analysis of TabNet masks, attributing feature importance at each decision step, thereby offering a more nuanced understanding of the prediction process. This is a significant deviation from traditional methods like SHAP values, enhancing understanding of feature interaction and decision-making processes. The study reveals that TabNet's focus on interpretability doesn't impede performance but rather facilitates it.

The research recommends future exploration of TabNet in related financial tasks like credit card or mortgage default predictions, which could provide new challenges and opportunities. It also suggests an examination of TabNet's performance across diverse geographies and economies, to uncover potential variations and expand its uses.

The development of advanced and interpretable models like TabNet holds potential to transform the financial industry by providing a clearer understanding of credit risks. The study further encourages research into more efficient methods for high interpretability and better handling of imbalanced datasets, both being current challenges in the field.

In sum, the thesis illuminates the promise of TabNet as an interpretable and high-performing model for loan default prediction, contributing to the ongoing discourse on balancing performance and interpretability in machine learning.

References

[1] Sercan O Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *arXiv preprint arXiv:1908.07442*, 2019.

[2] Giuseppe Cascarino, Mirko Moscatelli, and Fabio Parlapiano. Explainable artificial intelligence: interpreting default forecasting models based on machine learning. 2022.

[3] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774, 2017.

[4] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models. 2020.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.