Evaluating the Contribution of Open Banking Data to Credit Scoring Performance in the Spanish and French Markets

[1, 2]Francisco Mendonca
francisco.mendonca@revolut.com

[1, 2]Javier Ocariz Gallego
javier.ocariz@revolut.com

[1]Revolut Ltd
[2]Global Credit Management Data Science

Abstract - This paper investigates the discriminatory power of internal Revolut transaction data for developing a machine learning model to predict customers probability of default (PD). A primary challenge in credit modeling arises from data gaps, where specific variables are omitted by design during the underwriting process. This paper addresses a specific manifestation of this data heterogeneity: the absence of open banking data for customers who proceed through a particular underwriting funnel. To address this problem, a gradient boosting tree model is implemented utilizing forced splits, where the splitting criteria captures the existence of open banking data. This technique effectively trains two sub-models simultaneously: one for customers with open banking data and another for those without. This framework is used to train two credit scoring models for customers in the French and Spanish markets. The analysis demonstrates that the sub-models trained without open banking data relies more heavily on Revolut's own internal transaction data, and achieves similar discriminatory performance as the sub-models built on the full set of data, which includes open banking data. These findings suggest that open banking data may not offer a significant advantage in credit scoring, provided an alternative source of the customer's transactional data is available. This holds true even when the data originates from a non-primary account, thus evidencing that primary account usage offers limited information for creditworthiness.

1 INTRODUCTION

The integration of open banking data into credit scoring models has garnered increasing attention from financial institutions as they seek to enhance risk assessment capabilities ([1], [2]). This interest is driven by the potential of open banking data to provide a more granular and dynamic representation of borrowers' financial behavior, supplementing traditional sources of data. The ability to access transaction-level data enables lenders to infer patterns for dimensions such as income stability, discretionary spending, and financial resilience, which may be particularly valuable for assessing creditworthiness in populations with limited credit histories. However, there is the question of determining the extent to which open banking data contributes to model performance, particularly in markets where other sources of information are already well-developed. This study reveals that while open banking data is very useful for achieving high discriminatory power for recently onboarded customers with sparse interaction history, for more tenured customers, rich internal app usage data provides comparable predictive value, serving as an effective alternative.

This study examines the incremental predictive value of open banking data in credit risk models within the Spanish and French markets, when alternative sources of data around user spending habits exist. To quantify its contribution, we develop a credit scoring model for each market using the framework of LightGBM ([3]) coupled with a forced split framework. Both models are constructed using an identical customer pool, ensuring that any observed performance differences can be attributed solely to the presence or absence of open banking

data. The model performance was assessed using the ROC AUC. For both markets, we find that when Revolut's mobile app data is present, the contribution of open banking is very limited. These findings suggest that Revolut's app data is a strong competitor to open banking data, capturing similar customer behavior patterns.

This paper is organized as follows: Section 2 discusses model design and the forced split framework as well as the data sources; Section 3 presents a discussion of the results; Finally in section 4 the conclusions are discussed.

2 MODEL DESIGN

This paper addresses two distinct customer funnels: (1) one that captures open banking data and (2) a future customer funnel where such data is not captured. No historical data exists for the latter. Instead of developing and maintaining two separate models, which would increase operational complexity, a single, integrated model that relies on the concept of forced splits is proposed. The primary objective is to train a single predictive model capable of scoring customers in the aforementioned underwriting funnels. To achieve this, a data augmentation technique was employed. The original dataset, comprising personal loan applicants for which open banking data is captured, was duplicated to create a synthetic counterfactual cohort. Within this duplicated set, all features derived from open banking sources were set to null, thereby simulating the feature space of the future applicant funnel. The augmented dataset is used to train a machine-learning model using forced splits at the root of each tree.

Forced splits impose a specific, pre-defined architecture on the decision trees, acting as a constraint on the splitting criteria at the root node. This technique overrides the initial data-driven feature selection of the greedy algorithm, forcing the model to first and foremost partition the dataset into two mutually exclusive subgroups: (1) subjects for whom open banking data is available, and (2) subjects for whom it is absent. Once this primary segmentation is complete, the algorithm proceeds with its standard greedy optimization process recursively down each branch, using the full set of available features within that particular partition to determine subsequent splits. This forced-split architecture ensures that the model explicitly learns distinct predictive patterns tailored to the unique data characteristics of each cohort, thereby isolating the impact of the open banking data and enhancing model interpretability. Because open banking data is always missing on one of the subtree paths defined by the forced split, no open banking feature will be used in that subtree. In turn, this implies that the features and corresponding splitting strategy are allowed to differ across subtrees, resulting in different feature importance ranking and feature effects. This essentially results in two models nested into one. This approach, however, involves a well-defined trade-off between global interpretability and local predictive optimality. While forcing a split enhances the model's overall transparency, that specific split may not be the one the algorithm would have chosen to purely maximize information gain at that particular node. This can result in a marginal sacrifice of predictive power. The overarching goal is therefore to produce a usable model that satisfies crucial deployment constraints and interpretability requirements, rather than one that solely maximizes a given performance metric.

The datasets are split into train and test following a stratified sampling approach, with the test set being completely left out for validation purposes. Hyperparameters and feature selection are done using 5-fold cross-validation. The metric of interest to assess model performance is the AUC rank ordering metric. Feature importance and effect are evaluated through SHAP values ([4], [5]).

2.1 Forced Splits

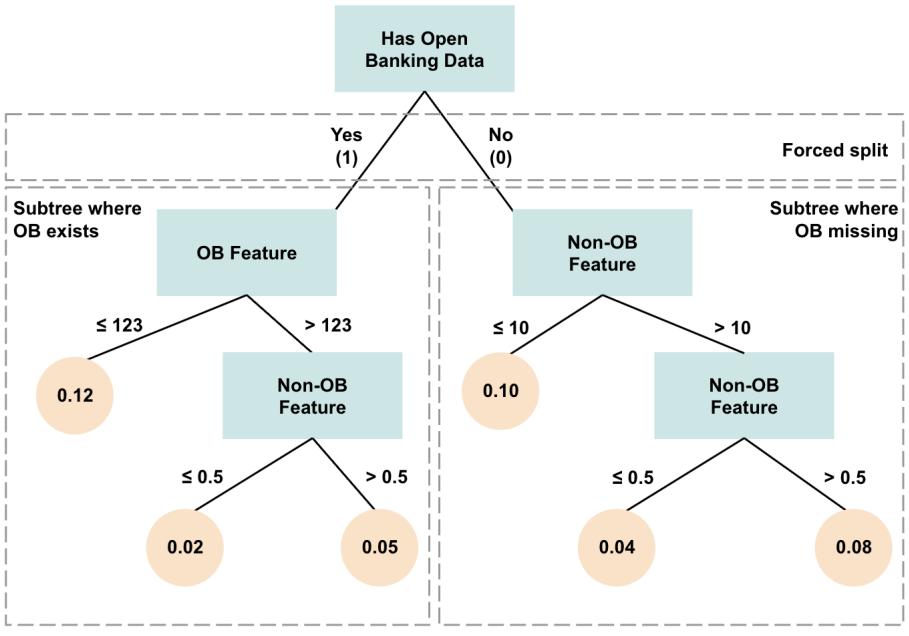
Using the LightGBM framework, forced splits are defined as parameter passed to the model constructor, as per example below in python programming.

```
import lightgbm as lgb
...
forced_split_json = {
    "feature": feature_list.index("has_open_banking_data"),
    "threshold": 0.0
}
## save json into forced_split_json.json

lgbm_parameters = {
    ...,
    forcedsplits_filename: "path/to/forced_split_json.json"
}
model = lgb.LGBMClassifier(**lgbm_parameters)
model.fit(X, y)
```

This setup forces the split using the feature *has_open_banking_data* right at the top of each tree, resulting in an unique tree path each customer can take depending on the existence of open banking. The figure below presents a diagram of this approach.

Figure 1: Example of a tree built using forced splits



2.2 Data

The data representative of personal loan and credit card applications for the French and Spanish markets was obtained from Revolut's internal datalake. The feature set is constructed from several distinct data blocks, encompassing sociodemographic variables, internal transactional activity captured via Revolut's mobile application, and metrics derived from open banking data. The target variable of interest in both markets is a binary indicator representing a delinquency event occurring within a pre-determined time horizon after loan origination. A notable distinction of the credit landscape in both France and Spain is the less developed role of traditional credit bureaus compared to other markets. In these countries, credit bureau data is not a conventional input for loan origination. Reflecting this, the models developed in this paper do not incorporate credit bureau data, placing a greater emphasis on alternative information sources like internal or open banking transactional data to assess creditworthiness.

The data for France contains 16626 observations of credit applications for personal loans and the corresponding credit outcome. The data is split into train and test. In the Spanish market, two products, personal loans and credit cards, were evaluated. Both utilized comparable acquisition funnels during the assessment period, suggesting similar AUC. The model training sample comprised 5056 personal loan observations. The test sample consisted of 3241 credit card observations. The use of different products for the train and test samples is acknowledged as a limitation due to the small sample size. However, this was deemed the best available option given the data constraints. This limitation is not expected to severely impact the results, as the AUC metric, which measures rank-ordering, is employed for evaluation. While delinquency ratios may differ between these products, it is expected that improvements in rank ordering will be consistent across them.

The features and corresponding sources that are used for each model are categorized in the table below:

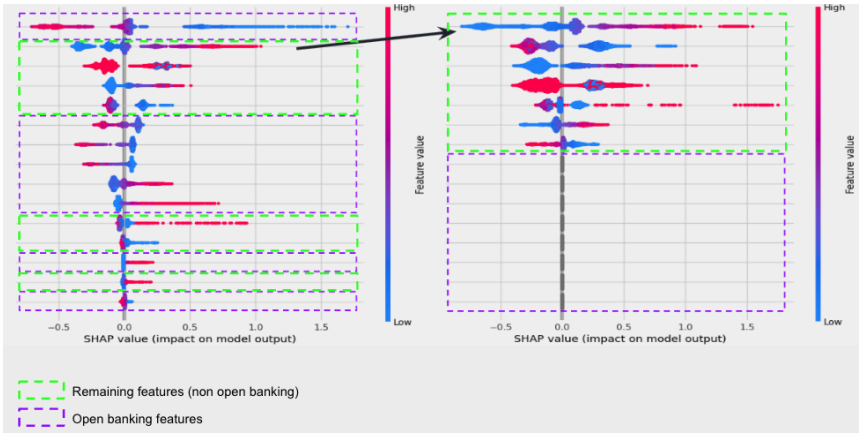
Source	**Description**
Application Data	Information provided by the customer at the time of onboarding, including basic demographic and economic declaration.
Past Credit History	Information about past credit applications and customer credit performance.
Internal App Usage	Information related to customers spending/income/savings/investment behavior, recorded in Revolut's mobile app.
Other Internal Data	Other type of data excluded from previous sources, existing within Revolut's database. Examples include device information, other declared information and similar attributes.
Open Banking Data	Data obtained after customer links external accounts. Reflects customer spending/income habits.

3 RESULTS

This section presents the results of the modeling framework applied separately to the French and Spanish markets. The models utilize the forced split methodology detailed previously to handle the different underwriting funnels. We compare feature effect across tree paths using SHAP values and measure the marginal impact of adding open banking data by comparing the AUC coefficients between tree paths.

Figures 2 and 3 provide a comparative analysis of feature utilization across the two distinct tree paths created by the forced split.Open banking features are denoted by purple boxes, while internal data features are shown in green. The SHAP plots clearly reveal that on the path where open banking data is available, these features rank highly in importance and contribute to the model's discriminatory power. Conversely, on the path where open banking data is absent by design, the model adapts and compensates by elevating the importance of the remaining features using a completely different splitting criteria to predict the outcome.

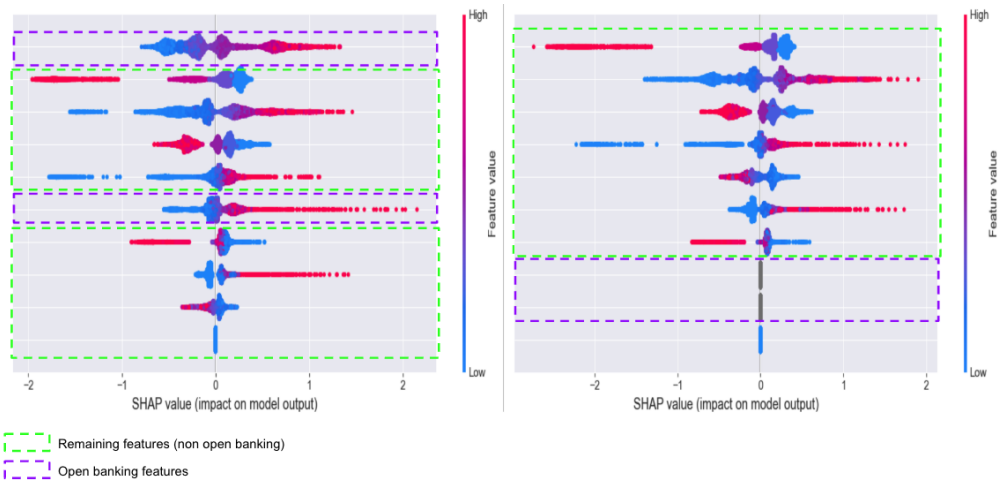
Figure 2: SHAP plots for the French model



(a) With Open Banking Data

(b) Without Open Banking Data

Figure 3: SHAP plots for the Spanish model



(a) With Open Banking Data

(b) Without Open Banking Data

Table 1 quantifies the marginal predictive lift offered by open banking data by presenting the absolute difference in the AUC between the two model paths created by the forced split. The results reveal that the sub-model trained without open banking data attains a comparable rank-ordering performance to its counterpart that leverages open-banking data. This indicates that the internal transactional data serves as a powerful standalone predictor of credit risk. While the inclusion of open banking data does yield a modest but consistently positive uplift in AUC, the magnitude of this improvement is marginal. This finding suggests that while open banking data is beneficial, it may offer diminishing returns when another rich source of customer transactional data is already present. The strong performance of the non-open banking path underscores the high predictive value inherent in the internal app usage and transaction data, which effectively functions as a robust proxy for the financial behaviors typically captured through open banking.

Table 1: Impact of including Open Banking data on model performance

Market	**Product**	**Train**	**Test**
France	**Personal Loans**	2.78 p.p.	3.26 p.p.
Spain	**Personal Loans**	4.00 p.p.	x
	Credit Cards	x	0.50 p.p.

4 CONCLUSIONS

This paper proposes a methodological framework for developing credit scoring models that can accommodate heterogeneous underwriting funnels, particularly where certain data sources are systematically missing. A unified model is constructed using a forced splits approach, with the primary splitting criterion being the availability of open banking data. The forced split methodology serves as a structural constraint on the model, that fundamentally shapes the learning process, yielding two primary effects. The most significant outcome is the creation of distinct, independent modeling paths within a single, unified framework. By partitioning the dataset at the root of each tree based on the availability of open banking data, the model is compeled to develop two specialized "sub-models." One is tailored to the data-rich environment of customers in the funnel capturing open banking data, while the other is specifically adapted for the data-scarce funnel, which relies more heavily on internal data. The framework is applied to the French and Spanish markets, and the analysis demonstrates that the resulting models maintain strong comparative rank ordering performance even in the absence of open banking data. SHAP analysis reveals a compensatory mechanism: when open banking data is not present, the model adapts by utilizing the remaining internal features differently to maintain its predictive power. The results therefore indicate that internal transaction data, such as that captured within Revolut's mobile application, serves as a robust and suitable alternative to open banking data for credit risk assessment.

REFERENCES

[1] L. O. Hjelkrem, P. Eilif de Lange, and E. Nesset, "The value of open banking data for application credit scoring: Case study of a norwegian bank," *Journal of Risk and Financial Management*, vol. 15, p. 597, 15 Dec. 2022. DOI: 10.3390/jrfm15120597.

[2] L. O. Hjelkrem and P. Eilif de Lange, "Explaining deep learning models for credit scoring with shap: A case study using open banking data," *Journal of Risk and Financial Management*, vol. 16, p. 221, 16 Apr. 2023. DOI: 10.3390/jrfm16040221.

[3] G. Kel, Q. Meng, T. Finley, *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, 2017.

[4] S. M. Lundberg and S.-I. Lee, "An unexpected unity among methods for interpreting model predictions," 2016.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, pp. 4765–74, 30 2017.