modulos

Revolutionizing Retail Credit Risk Management

Why Banks Need a Data-Centric AI Approach in 2023 and Beyond

Author: Elena Maran

Date: 10.03.2023

modulos.ai

Abstract

This whitepaper discusses **the comparison between a statistical methodology and the Data-Centric AI approach used by Modulos in retail credit decisions.**
The analysis is carried out with two datasets, an error free one, ML ready, and the other closer to real-world data quality with missing values, outliers, and incorrect information.

The results show that the **Data-Centric AI** methodology delivers **net superior results in terms of time and cost-saving, profit increase, losses reduction, and streamlining of the process.**

The **Data-Centric AI** methodology is able to deliver **an increase in profits of up to 35%, while reducing losses from insolvencies by up to 8% with respect to the statistical approach**. In addition, the final model results are achieved in a tenth of the time required by the statistical approach.

Both approaches are analyzed in detail, highlighting their strengths and weaknesses, and a particular focus is put on the **importance of data quality as the main driver of superior performance.**

Challenges and Opportunities in Retail Credit Business

The financial industry is facing unprecedented market conditions which pose a critical challenge to the sustainability of its retail credit business.

Extended periods of low-interest rates, and the entry of new players from the FinTech industry, with lower regulatory constraints, using state-of-the-art technologies and with a leaner structure that allows for faster decision-making, are all factors putting traditional financial institutions under strain and calling for immediate action.

Unprecedented market conditions are challenging the sustainability of the retail credit business.

Witnessing declining net interest margins, traditional retail banks are prompted to seek ways to **expand their lending base without compromising on credit quality.**

One approach is to move away from deterministic, rule-based methods of credit assessment and adopt more sophisticated statistical methodologies that provide a probability of default for each individual applicant.

However, even these methods still have limitations, and **the industry is now looking at the potential of Artificial Intelligence (AI) to revolutionize credit risk assessment.**

This whitepaper explores how Modulos' Data-Centric AI can complement statistical methods to provide a more accurate, efficient, and effective solution for credit risk assessments.

The analysis will be carried out on a specific problem of creditworthiness determination, where 1,000 customer observations and 21 variables have been taken into account.

The goal is to **determine whether an existing or a new retail customer of a financial institution is eligible for a personal loan** by classifying them as an acceptable or unacceptable credit risk. Of the 1,000 observations available, 44% represent defaulting applicants, which are considered unacceptable credit risks.

1

Personal Information

Including age, sex, employment status, marital status

2

Loan details

Including amount, purpose, potential credit mitigants - e.g. guarantor

3

Past credit history and relationship with the bank

Other credit exposures, repaymenthisto-ry, tenure with the bank

The available information in the form of a dataset pertains to three main sections. The first contains **personal information**, such as age, sex, employment status, and marital status. The second refers to the specific details of the **loan exposure**, including the amount, purpose, and existence of credit mitigations. The last pertains to **past credit history and the relationship with the bank**, including other credit exposures, repayment history, and tenure with the bank.

In this context, a performance comparison of statistical methods versus Data-Centric AI is carried out to **determine which approach is more effective in credit risk assessment**.

A step-by-step comparison of statistical methods versus Modulos' Data-Centric AI is carried out to determine which approach is more effective in credit risk assessment.

We will follow a step-by-step process to analyze the problem, compare the approaches, and evaluate their results. By the end of this whitepaper, readers will have a clear understanding of the limitations of statistical methods and the advantages of Data-Centric AI in credit risk assessment, with its potential to transform the financial industry.

Unpacking Credit Risk Assessment: A Detailed Comparison of Two Workflows

Statistical approach

At a high level, the statistical methodology relies on a progressive iteration where **variables are transformed and reduced** until a small number of variables are used in a logistic regression model to determine whether to lend money or not.

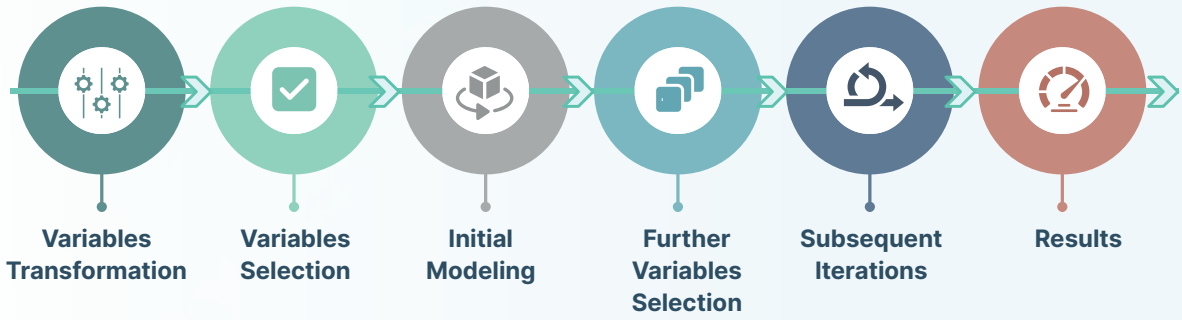
Data-Centric AI approach

On the other hand, the Data-Centric AI approach focuses on addressing potential data problems that can negatively impact model performance and enables a **feedback loop between data and model** to continuously improve performance.

The Data-Centric AI approach offers a more streamlined, objective, and scalable solution to credit risk assessment.

The statistical methodology is a time-intensive, manual process that is often subjective because the variables selection is heavily influenced by the statistical criteria used. In contrast, **Modulos' Data-Centric AI workflow is fast, efficient, and prioritizes data quality as the primary driver of model performance**. Modulos approach eliminates the need for manual feature selection and reduces the risk of missing important information that can impact the credit decision. Ultimately, the Data-Centric AI approach offers a more streamlined, objective, and scalable solution to credit risk assessment.

Statistical approach



Data-Centric AI approach

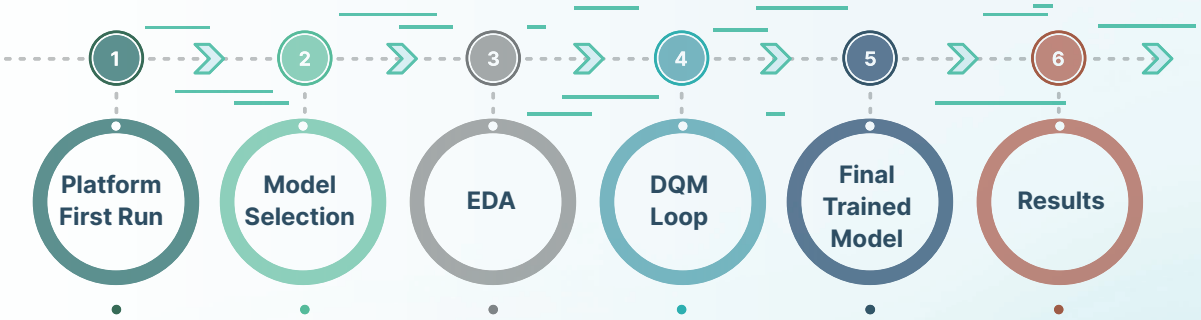


Figure 1. Statistical approach vs Data-Centric AI approach

Phase 1:
Pre-processing versus initial results

Statistical approach

While the **statistical approach involves a preliminary step of variables transformation**, which includes categorization and aggregation, Modulos' Data-Centric AI methodology begins with an initial run of the platform to identify available models and assess their performance scores.

Specifically, the statistical approach transforms both categorical and continuous numerical variables into categories, and aggregates categories with similar behavior, the so-called data binning. For instance, in the case of checking accounts, individuals with no money and those with more than €200 are aggregated together. This approach addresses issues with data quality, such as missing values and outliers, by fitting them into predetermined categories.

Feature: Checking account status										KS: 24.4%
Categories	**0**	**1**	**Total**	**%Flag**	**% Flag0**	**% Flag1**	**% tot**			
none	143	182	325	56.0%	31.2%	53.2%	40.6%	31.2%	53.2%	22.0%
x>= 200 Euros	25	27	52	51.9%	5.5%	7.9%	6.5%	36.7%	61.1%	24.4%
0<=x<200 Euros	130	70	200	35.0%	28.4%	20.5%	25.0%	65.1%	81.6%	16.5%
x<0 Euros	160	63	223	28.3%	34.9%	18.4%	27.9%	100.0%	100.0%	0.0%
Total	458	342	800	42.8%	100.0%	100.0%	100.0%			

Transformed - Checking account status										KS: 24.4%
Categories	**0**	**1**	**Total**	**%Flag**	**% Flag0**	**% Flag1**	**% tot**			
none + x>= 200 Euros	168	209	377	55.4%	36.7%	61.1%	47.1%	36.7%	61.1%	24.4%
0<=x<200 Euros	130	70	200	35.0%	28.4%	20.5%	25.0%	65.1%	81.6%	16.5%
x<0 Euros	160	63	223	28.3%	34.9%	18.4%	27.9%	100.0%	100.0%	0.0%
Total	458	342	800	42.8%	100.0%	100.0%	100.0%			

Figure 2. Variables transformation in the statistical methodology

Whilst the statistical methododology starts with variables pre-processing through data binning, Modulos methodology evaluates the performance of several models, given the available data and the problem at hand.

Data-Centric AI approach

In contrast, the **Data-Centric AI methodology starts by running the platform to identify available models and assess their performance scores**. This allows for a more dynamic and adaptive approach, as the performance of different models can be evaluated and updated based on the available data.

Figure 3: First run of the platform in Modulos Data-Centric AI

Phase 2:
Variables versus model selection

Statistical approach

Continuing our comparison of the two credit risk modeling workflows, the **second step of the statistical approach involves the variables selection**. A Kologorov-Smirnov (KS)[1] test is performed on all available variables to select those that are statistically relevant and will be used for modeling. This is itself a very time-intensive process as a KS score needs to be computed for every category within a variable, to determine how much it explains the credit decision. The KS score of the overall variable is the highest score of its categories.

Variables	KS Construction	KS Validation	Logistic Step 1	Analysis
checkings_account_statusNew	24.0%	23.0%	Y	Strong
duration_monthsNew	41.0%	7.0%	N	-
credit_historyNew	20.0%	19.0%	Y	Strong
purposeNew	15.0%	17.0%	Y	Strong
credit_amountNew	32.0%	48.0%	Y	Strong
savings_account_or_bondsNew	9.0%	10.0%	Y	Good
present_employment_sinceNew	13.0%	9.0%	Y	Good
installment_rate_percentage_disp	3.0%	4.0%	N	-
gender	10.0%	17.0%	Y	Strong
marital_status	6.0%	7.0%	Y	Week
debtors_guarantors	2.0%	0.0%	N	-

Figure 4: The KS test for variables selection

The statistical approach focuses on variables selection using statistical tests, while Modulos' Data-Centric AI methodology leverages an Auto ML engine to identify all potential models suited to the credit use case.

[1] *The KS test is one of the statistical methods to assess the explanatory value of a feature. This analysis has focused on a combination of KS test and Information Values to reduce the number of variables used for the determination of credit worthiness*

Data-Centric AI approach

In contrast, **Modulos' Data-Centric AI methodology at this stage selects a model rather than individual variables.** An Auto ML engine identifies potential models, and the best-performing one is typically selected at this initial stage. However, if constraints exist, such as the need for explainability, the search can be restricted to more interpretable models such as three-based models or logistic regressions.

Overall, both approaches have different ways of selecting variables or models depending on the specific problem requirements. The statistical approach focuses on variable selection using statistical tests, while the Data-Centric AI methodology leverages an Auto ML engine to identify potential models.

Phase 3:
Logistic regression modelling versus Exploratory Data Analysis

Statistical approach

In the third step the **statistical methodology's focus shifts towards the modelling** of the credit risk problem, using the variables selected in the prior step **to run a logistic regression.**

Data-Centric AI approach

Modulos approach at this point starts an **Exploratory Data Analysis (EDA) to identify any potential data quality issue** before proceeding to the next step.

With Modulos, users have access to functionalities that allow them to easily visualize and comprehend where data quality problems may originate.

Data-Centric AI approach

For instance, as displayed in the image below, users can quickly see that “age” is a variable that has a high number of missing values, while “credit amounts” displays a high skew. With this understanding, the subsequent steps of the modeling process can focus on where the data quality problems originate from, allowing for more effective and efficient modeling.

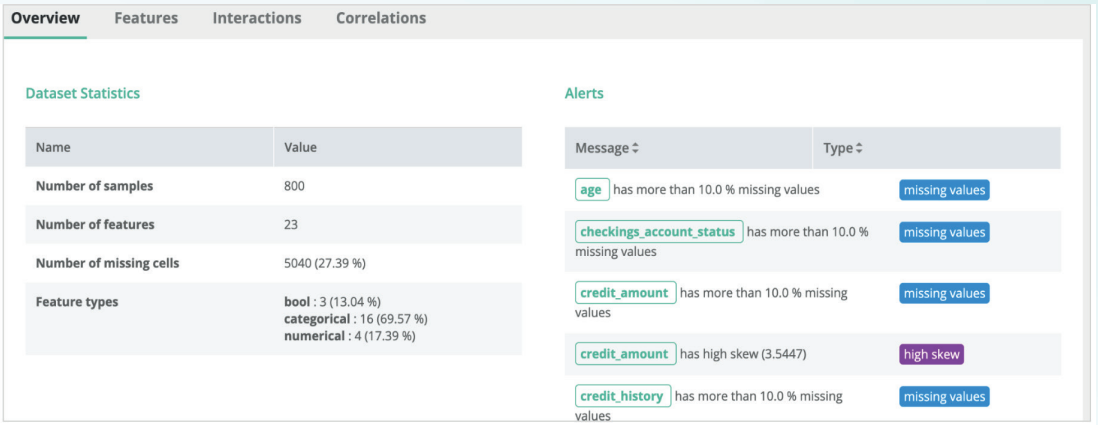


Figure 5: EDA with Modulos

Phase 4:
Further variables selection versus Data-Model feedback loop

Statistical approach

Moving on to the fourth step, **in the statistical approach** to credit risk modeling, **the focus is on continuing the iterative selection of the variables**. Using the results of the first logistic regression from the previous step, the pool of variables is further reduced by examining the statistical measure called Information Value for each of the variables. At this point, only seven features remain for use in subsequent iterations of the logistic regression.

Data-Centric AI approach

Modulos’ fourth step involves a **very innovative approach to improving the initial model performance through a Data-Model feedback loop**, whereby after improving the data the model is retrained and its accuracy is reassessed. With Modulos, users can easily identify sources of error, noise, and bias within their data that may limit model performance.

With Modulos, users can easily identify sources of error, noise, and bias within their data that may limit model performance.

The platform allows for the **identification of individual training samples that are detrimental to model performance**. These are the samples highlighted in orange in the below chart, which require further analysis, with a focus on the features that the previous EDA had revealed as being the most problematic ones.

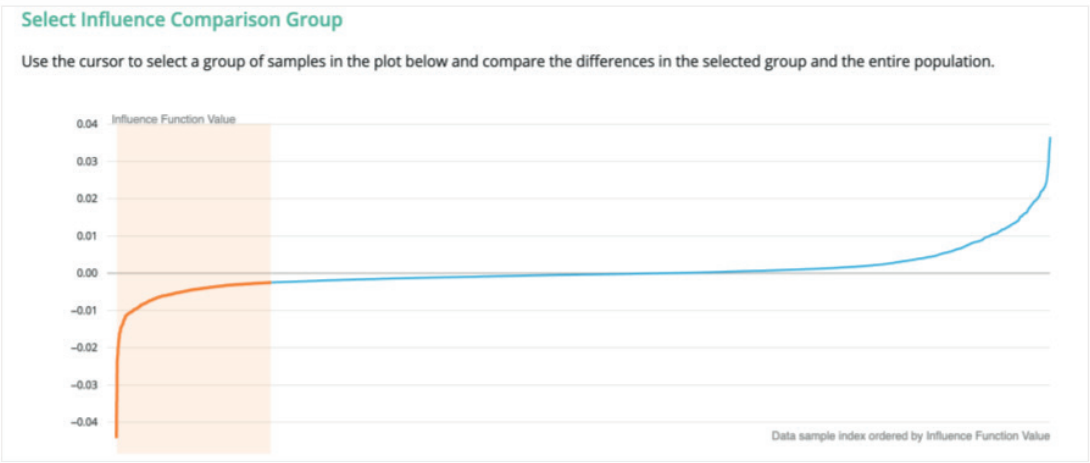


Figure 6: Training samples that impair model accuracy

By following the platform's recommendations and engaging in subsequent iterations of the data cleaning process followed by model retraining, users can progressively improve model performance.

Phase 5:
Subsequent iterations versus a selection of the final trained model version

Statistical approach

On the fifth and last step before coming to the results, **the statistical credit risk modeling approach continues its iterative variables selection.** This is achieved again by looking at the Information Value of the various variables and plugging them into different rounds of logistic regression. Additionally, different transformations are performed by aggregating categories with similar statistical behavior. This process continues until the final step, where typically four to six variables are left to explain the problem.

Data-Centric AI approach

In contrast, with Data-Centric AI, after the progressive iterations aimed at improving the model performance by retraining it with an ever-improved dataset, it is time to select the **final model version.** This is the version for which we are satisfied with the performance, in this case in terms of accuracy. Once the final model has been selected, it can be deployed and put into production.

The statistical methodology iterates on variables' selection to limit the number of explanatory features, whereas Data-Centric AI enables a Data-Model feedback loop.

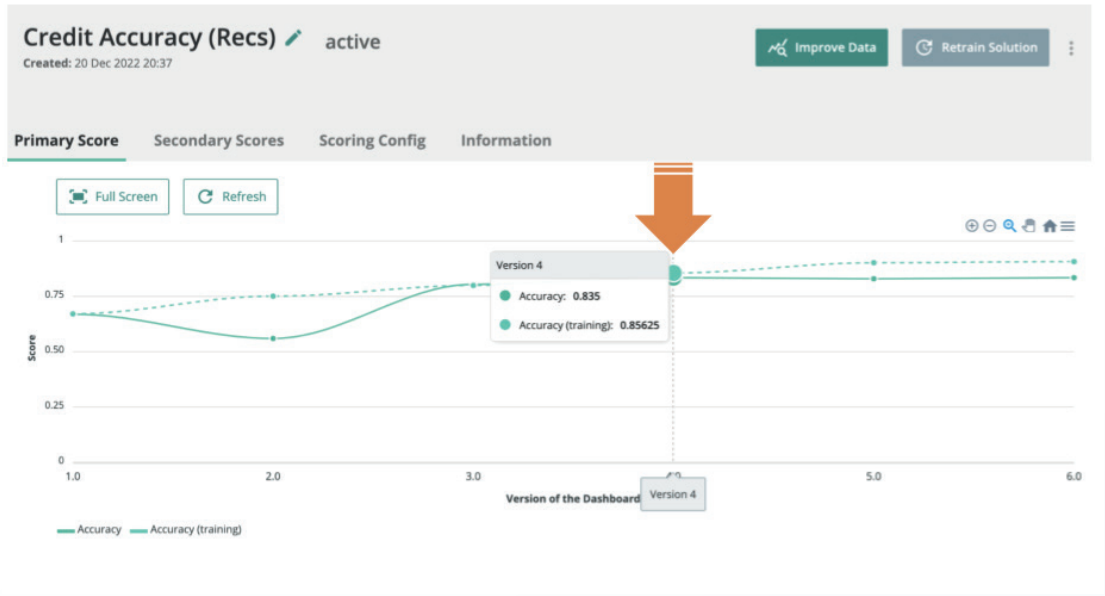


Figure 7: Iterative model performance improvement

Revolutionizing Retail Credit Risk Modeling: Unveiling Impressive Results with Data-Centric AI

The previous session of this whitepaper has provided a very detailed description of the critical differences between the two methodologies of retail credit risk modeling that have been compared: the statistical approach and Modulos' Data-Centric AI.

The analysis started by looking at the clean data set, which, although the least realistic, provides a great basis for comparison.

The results were nothing short of impressive - **Modulos' methodology delivered 10% higher accuracy**, allowed to increase the disbursement of performing loans by 21% - translating into higher profits - and reduced non-performing loans disbursed by 37% - translating into significantly lower losses.

Modulos' methodology delivered 10% higher accuracy, allowed to increase the disbursement of performing loans by 21% and reduced non-performing loans disbursed by 37%.

These results are captured by the confusion matrix, where the differences in true positives and false positives are measured - the most crucial measures in determining profits and losses.

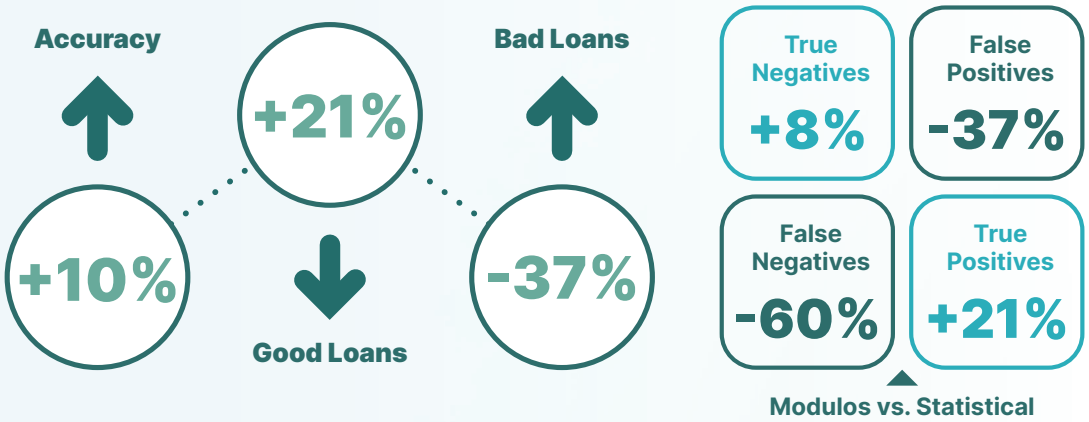


Figure 8: Results comparison on a clean dataset

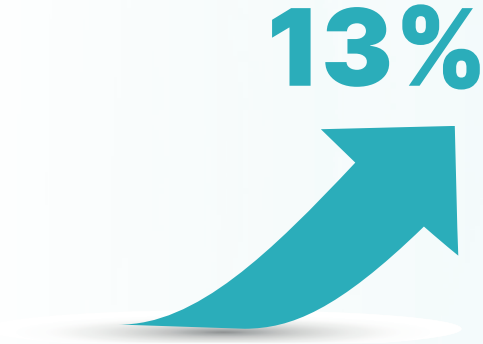
Figure 9: Business case with clean dataset

By making a few reasonable assumptions about net interest margin, and loss given default, it can be seen how Modulos' methodology would result in an additional $470 in additional profits and $1350 in losses saved per every 100 applications of creditworthy and insolvent applicants respectively, with an average amount lent of $1000. These results are just the tip of the iceberg - for financial institutions with thousands of clients and lending larger notional amounts, the benefits of implementing Modulos methodology can be substantial.

A further comparison of the results of Modulos' methodology with the statistical approach was performed, this time using the same logistic regression mode.

Even when constrained to use the same model, Modulos' methodology still delivered an accuracy increase of 7%, an increase of 21% in "good" loans, and a reduction of 5% in "bad" loans. The crucial difference lies in the selection of variables - **Modulos is not restricted to focus on a small pool of variables, thus being able to capture complex relationships**. A variable like "duration of the relationship", which is a crucial variable in the Data-Centric AI workflow, is instead ruled out by the statistical modeling on the very initial stage.

Moving on to the more realistic dirty data set, **the Data-Centric AI approach saw an incredible 13% increase in accuracy**, 35% more "good" loans, and a 6% reduction in "bad" loans, compared to the statistical approach.



When performing the analysis on a lower quality dataset by following Modulos' platform recommendations, the accuracy of the model was improved from 67% to 83.5%

Modulos' platform started with an accuracy of 67% and, following the recommendations for data cleansing, improved to an impressive 83.5% accuracy level. **Data quality played a particular important role in the identification of applicants** which would not default after the loan disbursement, as evidenced by the impressive 35% increase in good loans.

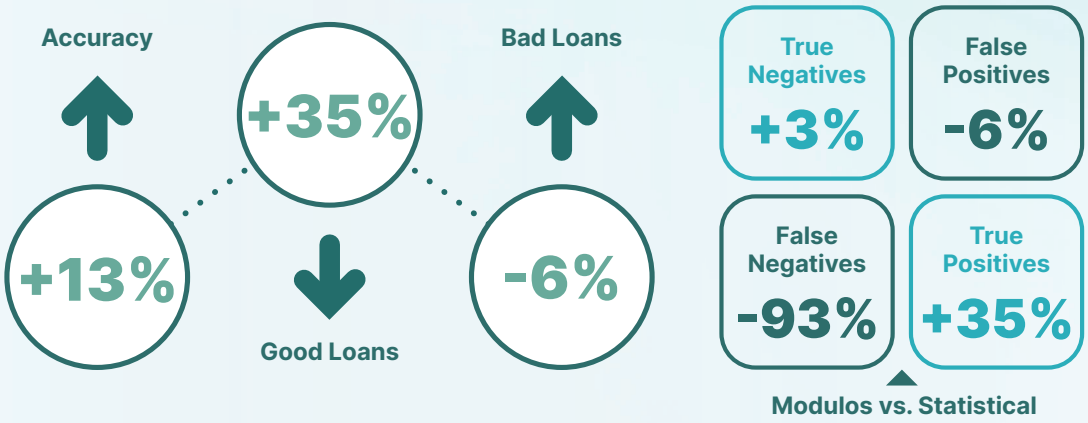


Figure 10: Results comparison on a dirty dataset

Using the same assumptions as before, **Modulos methodology would allow for an additional 25 loans per every 100 good applications, translating to $750 in additional profits.**

The numbers are staggering, and the benefits of implementing the Data-Centric AI methodology are evident.

Data Quality: The Key to Superior Accuracy in Retail Credit Risk Modeling

Data quality is a crucial factor in the performance of any predictive model, and the handling of data quality is where the Modulos platform truly excels.

In the statistical credit risk modeling, data quality is typically addressed only during the pre-processing stage, where outliers or missing values are identified and fit into one of the given variable categories. However, this approach does not always deal with all the issues in the data. If a variable has a substantial presence of missing values and the distribution is critical to explaining the outcome, it will be given a high statistical KS score and used for the modelling even if it can degrade the model performance.

Modulos allows to swiftly address errors, biases, and noises in data, resulting in significant improvements in model performance.

Modulos, on the other hand, offers a more detailed approach to data quality, allowing to identify the specific samples in a dataset that are negatively impacting performance. By examining the distribution of different variables, such as “age” or “credit amount”, it is possible to pinpoint where the problems with the data are concentrated. **Modulos allows to analyze the contribution of individual data samples to the model’s accuracy and to swiftly address errors, biases, and noises**, resulting in significant improvements in performance. **With Modulos, one can be confident in the quality of the data and the accuracy of the modeling.**

Conclusions

In today's rapidly evolving financial landscape, traditional banks are facing unprecedented challenges to their retail credit business. With lower interest rates and new fintech competitors using advanced technologies to gain a competitive edge, it is imperative that banks adapt quickly to these market conditions.

This is where Modulos comes in, providing an innovative and sophisticated solution to credit risk assessment that puts traditional rule-based methods to shame and performs better than the more sophisticated statical methodology that has been analysed in this whitepaper. By leveraging, on its revolutionary methodology of dealing with Data Quality issues, **Modulos offers unparalleled accuracy and speed, allowing banks to expand their lending base without compromising on credit quality.**

With Modulos, banks can quickly identify potential sources of error, noise and bias in their datasets, and ultimately **achieve superior performance** in their credit risk models.

Modulos results speak for themselves - **with the Data-Centric AI methodology significant improvements in model accuracy have been achieved, reducing false positives and false negatives**, and ultimately mitigating risk and increasing profitability in just a fraction of the time required by the statistical approach.

If you are looking to stay ahead in today's competitive financial landscape, Modulos is the solution you need.