A User-Centered Approach to Create Realistic Datasets for AI. Case Study: Creditworthiness in the Banking Sector

Francesca Zampino[1], Antonella Longo[1] and Marco Zappatore[1]
[1]*University of Salento, via Lecce-Monteroni, 73100 Lecce (LE), Italy*

Abstract

Nowadays businesses are evolving, as new digital tools ensure greater efficiency of their information systems. Decision-making and strategic processes can benefit from innovation opportunities such as Machine Learning. The main issue encountered in Artificial Intelligence applications, is that data can be not available or unsuitable for the case of study. This paper proposes the solution for this problem, by generating simulated data for AI. The case of study is creditworthiness in the banking sector; a loan is considered the main source of income for the banking sector, as well as the main source of risk. Consequently, the evaluation of creditworthiness is a key activity both for banks and for customers. To address this need, we propose a solution tailored to lenders to evaluate credit applications and to customers to be aware of behaviors that can reduce their credit score.
The approach proposed in this paper aims at realizing realistic datasets for Artificial Intelligence (named IDEA) to meet specific business needs, and to respect users' requests. An analysis of the current literature and methods for the evolution of conceptual models will be conducted, through pre-existing datasets. The proposed approach draws from and extends such literature. The intended application is to adopt this approach in the banking sector for considering the creditworthiness of customers who have entered into financial relationships. Therefore, the envisaged use case is to forecast the probability of borrowers going bankrupt. The paper defines the approach applied to specific financial datasets for the use case. Moreover, a validation of datasets is done, thanks to the Data Quality Index, before applying IDEA to predict credit solvency.

Keywords

Artificial Intelligence, realistic datasets, user-centered approach, prediction, creditworthiness

1. Introduction

The loan is a core business for the banking sector, as well as the main source of financial risk for banks. European data show that the loan is the most widely used financing instrument for small and medium-sized enterprises. The situation in which an asset causes high risks due to the inability of a borrower to return the loan within the agreed time; is called a "Credit Risk"[3]. A borrower's creditworthiness was based on a numerical value, a score named "Credit score". In general, this value helps authorities calculate the likelihood that a borrower will return the loan within the designated time. Creditworthiness means the ability of a debtor (in this case, a financial intermediary) to repay its debts on maturity, based on credit history or payment history.

Recently, researchers and banks have chosen training classifiers based on various machine learning and deep learning algorithms to automatically predict an applicant's credit score based on its credit history and other historical data [3]. For example, we can calculate the future score of the credit score or the probability of default, before issuing a loan. In order to reach our goals, the process is divided in the steps now explained.

ITADATA2022: The 1st Italian Conference on Big Data and Data Science, September 20-21 2022, Milan, Italy

EMAIL: francesca.zampino@studenti.unisalento.it (A. 1); antonella.longo@unisalento.it (A. 2); marcosalvatore.zappatore@unisalento.it (A. 3)
ORCID: 0000-0002-6902-0160 (A. 2); 0000-0002-8277-9390 (A. 3)

- First, we proceed to the research and selection of scientific papers targeting the same goal of the study: this allowed us identifying the financial variables of interest and the corresponding datasets.

  Understanding the elements that identify the context is useful for the implementation of a consistent database model. The analysis process starts from the choice of models and related variables, considered useful to describe the banking context. In general, the literature is characterized by datasets that identify a loan as a reference entity, with attention to the credit history of the applicants.
- Secondly, it is important to verify the applicability and value of the model applied to different bank cases.
- The model must be validated to evaluate the quality of datasets.

The paper is organized as follows:

Section 1 introduces the literature analysis we performed to ground the proposed approach, which is described in details in Section 2. Section 3 discusses a dataset validation, based on a dedicated Data Quality Index, along with the achieved results. Conclusions are drawn in Section 4.

2. State of art

This section presents an analysis on scientific papers chosen as a standard reference on top of which the proposed IDEA approach is grounded. This analysis has been carried out to evaluate the widespread models for traditional and innovative banking realities. Therefore, the literature review is the starting point for our research. We focused on two main topic typologies:

- Commercial banking
- Peer to Peer lending

The first category of identified papers includes traditional bank loans.[1][2][3][4].

Instead, the second category refers to a kind of financial innovation, Peer to Peer lending, a loan between individuals, granted without traditional financial intermediation.[5][6][7]. The analyzed papers present an analysis of the banking sector and choose to apply SEMMA as a data mining design model. The SEMMA method is more useful than the alternative model, CRISP DM because SEMMA pays attention to user requests, asked by our study.

SEMMA [8] is the multi-stage method applied by the papers analyzed.

1. SAMPLE: Firstly, the goal is to identify a representative model for the population. The process of collecting data from the whole population is a very difficult task, so SEMMA offers the opportunity of using a sample of population data for the development of the model.
2. EXPLORE: The next step of the SEMMA methodology is data review.
3. MODIFY: The main tasks related to data modification are the conversion of data types and the management of missing values.
4. MODEL: In this step of SEMMA, several algorithms and mining techniques are applied to develop the proposed model. The purpose of this step is to identify the hidden and meaningful information from the pre-processed data set. Among the algorithms used, Decision Tree, Logistic Regression and Neural Network.
5. ACCESS: Once the implementation and validation of the model has been satisfied by all the proposed techniques, the test data is incorporated into each model, for the loan approval prediction.

The SEMMA methodology is widely used in the literature and can be compared with two other tools for machine learning models: CRISP-DM and KDD [9]. SEMMA and CRISP-DM are an evolution of KDD (1996).

The CRISP-DM standard was published in its first version in 1999 in Brussels and it is composed of 6 main stages, which can be added at the end. The steps are:

1. Business Understanding: understanding business problems.

2. Data Understanding: understanding data is fundamental to understand how data and analysis can solve the problem of the previous phase.
3. Data Preparation: this is the data cleaning and review phase.
4. Modeling: it is the choice of the algorithm suitable for the use case.
5. Evaluation: an evaluation of the outputs: it will be possible to use a part of the data, the test ones, to compare the results of the model with the real ones.
6. Deployment: the model will have to go into production, that is, it will have to be used on a large scale.

The SEMMA method allows the development of an application domain for end-user goals; in the SEMMA Sample phase, data cannot be sampled unless there is an understanding of all business needs.

The approach developed by this study stems from the so-called "business goals", business purposes to be defined upstream, as a reference point for the model to be developed. It is better represented by the SEMMA method, for the reasons explained. The aim is to create a suitable model to meet the needs of users. 1. During this study, the data samples were selected from Kaggle Repository.

3. The proposed approach: IDEA

In this Section, we will discuss IDEA, (real**I**stic **D**atas**E**t for **A**rtificial intelligence) a systematic model approach designed to respond to business needs, so that the expectations of target users can be addressed properly. IDEA is an extension of the SEMMA model, discussed in Section 2. We will identify the requirements of companies and users in order to carry out an in-depth search on existing data repositories.

On the one hand we will evaluate different data sources to develop an optimal combination of variables since our purpose is to maximize strategic business benefits.

On the other hand, the research and choice of datasets will be followed by the development of a conceptual model as a graphic representation of the context. A conceptual model can explain the main entities and relationships between them. It will be populated by considering datasets suitable to represent the analysis scenario.

IDEA aims to identify the process of borrowing activities, basically in the form of a loan, whose borrowers are individuals from Italian regions of North, South and Center, aged 25 years and over.

In order to apply the proposed approach, an open dataset for commercial banking was selected, because granting loans to private individuals is the core business of a commercial bank. The dataset choice is motivated by the purpose to provide a basic model for banking institutions which can also be turned into a more complex model, such as P2P lending. At the same time, it is also useful for small traditional banking companies.

Our aim is to demonstrate the applicability of the proposed IDEA approach to all banking realities, small and medium-sized enterprises or companies with high turnover. This means that the model can be used for traditional and innovative banks, because IDEA is presented as a standard model with traditional features, but it can be changed thanks to innovative variables, for example P2P lending ones. This model does not focus on the number of variables, but it is characterized by functional attributes to identify entities and context. A limited number of attributes is due to the choice of creating a model for small banks which can be developed through other variables to become a large bank model. The process can be applied from micro to macro realities.

The model is provided as a general application guideline that can be adapted to the banking reality that decides to apply it. The underlying entities of any financial relationship are customer, loan, and bank.

Moreover, Figure 1 shows a real estate entity, linked to the loan, through a non-compulsory relationship: the asset can be a guarantee for the financial relationship. Although the model represents financial deals, it was chosen to specify the optional loan guarantee to ensure the assessment of the solvency of creditors and Probability of Default. Our dataset can be different because it does not cause the operational problems of data normalization, about null values or duplicated data. Our model also allows you to identify the relationships between customer, loan and bank. IDEA defines each key attribute, while, for Kaggle datasets, we should integrate the key data, by generating key attributes randomly.

In fact, the use case of the model is forecasting the evolution of a credit portfolio, in terms of its financial reliability. IDEA can be, therefore, critical to define a bank strategy.

The research about relevant banking datasets is also characterized by an evaluation of different online sources and data repositories currently available today. Among these Kaggle[2], Dataport[3], World Bank[4], World Economic Forum[5], Towards Data Science[6] and Data. World [7]were considered. These datasets are open, but it is important to carry out an initial verification and selection of the attributes, understandable and, at the same time, consistent with the model. After a careful analysis, the most consistent repository was considered Kaggle, for the availability of all variables. As explained previously, IDEA focuses on a limited number of attributes that identify a small, medium, or large bank. The three main variables we have considered are loans, customers and guarantees that were found in the Kaggle datasets deemed suitable for the model. This source is the best one to represent a development from micro to macro realities.

The proposed methodology aims at identifying the main entities (clients, loans, real estate and guarantees) and corresponding attributes to develop a suitable Entity-Relationship conceptual model and then build a physical relational database. It is important to define the reference entity as a loan, identified by specific attributes properly related to other entities such as the borrower - client. Literature analysis allowed us to identify the IDEA main attributes. Elements in most of the articles are the following:

- Id - loan
- Id - borrower
- Sex
- Personal data
- Education
- Income of the borrower (main borrower)
- Income of the second debtor
- Amount of financing
- Duration of financing in months
- Credit history
- State of financing
- Interest rate
- Spread on interest rate
- Installment
- Date of issue of the loan
- Purpose
- Default of the loan (1 = defaulting borrower; 0 = fulfilling borrower)
- Card code (YES / NO)
- Credit score
- Year of birth
- Level of credit
- Age
- Up front charges

In addition, these variables are related to the opportunity that a loan is secured by real estate.

- Type of warranty

[2] https://www.kaggle.com/
[3] https://ieee-dataport.org/
[4] https://www.worldbank.org/en/home
[5] https://www.weforum.org/
[6] https://towardsdatascience.com/
[7] https://data.world/

- Type of building
- Amount of property evaluation

These elements were used to build a database, whose conceptual model is partially shown in Figure 1. Since we have created the model to make the estimate of solvency, we proceeded with its preliminary validation, before applying it to a case study. In the next Section, the validation of IDEA is discussed.

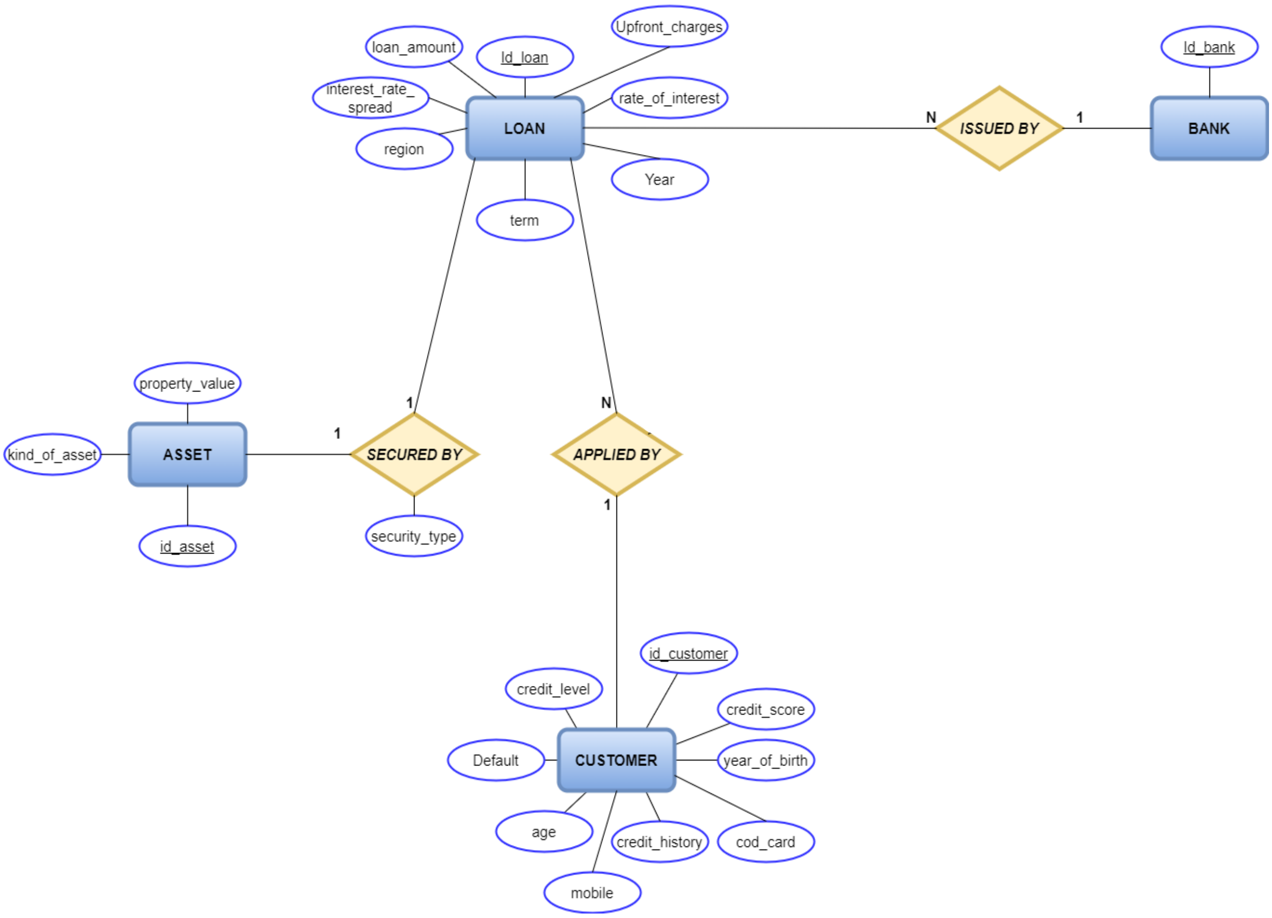


Figure 1: Entity-Relationship diagram

4. Data set validation: the Data Quality Index (DQI)

This section addresses the dataset quality for the IDEA approach [10]. We define an assessment parameter (the Data Quality Index), characterized by the following weighted metrics (each weight indicates the importance we attribute to its impact on data quality):

- Accuracy (20%)
- Completeness (20%)
- Consistency (20%)

- Uniqueness (20%)
- Validity (10%)
- Integrity (5%)
- Timeliness (5%).

These parameters can be estimated by analyzing dataset features. This analysis can be done using Python tools. A dedicated function (i.e., "Pandas Profiling") from Pandas, the widely used Python data management library, was used for the evaluation of these metrics. After analyzing the dataset attributes, each metric is evaluated from 1 to 5. At the end these results contribute to estimate the Data Quality Index, which is a weighted sum of each parameter. In Table 1 and Table 2 an explanation of values from 1 to 5 is shown.

Table 1

Metrics values

Metrics	Questions	
Accuracy	1.	**Percentage of data with no misspellings**
	1	0%
	2	40%
	3	60%
	4	80%
	5	100%
Completeness	1.	**Percentage of missing cells**
	1	>20%
	2	>10%
	3	>5%
	4	>2.5%
	5	0%
Consistency	1.	**Correlation between attributes**
	1	0.2
	2	0.4
	3	0.6
	4	0.8
	5	1
Uniqueness	1.	**Percentage of duplications**
	1	>20%
	2	>10%
	3	>5%
	4	>2.5%
	5	0%
Validity	1.	**Amount of data that makes the dataset representative of reality**
	1	250
	2	500
	3	1000
	4	>100000
	5	>200000
Integrity	1.	**Percentage of empty database fields**
	1	>20%
	2	>10%
	3	>5%
	4	>2.5%
	5	0%
Timeliness	1.	**Is data updated?**
	1	<1990
	2	>1990
	3	>2000

	4	>2010
	5	>2020

Table 2
Metrics values

Metrics	Questions	
Accuracy	2.	**Source reliability**
	1	Private source
	2	Chargeable source
	3	Auto-realization source
	4	Public private source
	5	Public source
Completeness	2.	**Percentage of missing values for each field**
	1	>20%
	2	>10%
	3	>5%
	4	>2.5%
	5	0%
Consistency	2.	**Correlation between fields**
	1	0.2
	2	0.4
	3	0.6
	4	0.8
	5	1
Uniqueness	2.	**Percentage of duplications for each field**
	1	>20%
	2	>10%
	3	>5%
	4	>2.5%
	5	0%
Validity		**Amount of data to make dataset reliable**
	1	250
	2	500
	3	1000
	4	>100000
	5	>200000
Integrity	2.	**Percentage of correct values**
	1	100%
	2	>80%
	3	>60%
	4	>40%
	5	0%
Timeliness	2.	**Data update frequency**
	1	0
	2	20 years
	3	10 years
	4	5 years
	5	<5 years

The evaluation is based on these questions about datasets:

- Accuracy:

1. Are there any spelling errors in the data names?
2. Do data accurately represent the "real world" values they are supposed to detect?

- Completeness:

1. Are there data values with null elements for the entire dataset?
2. Are there data values with null elements per field?

- Consistency:

1. Are data presented in a similar or compatible format?
2. Are there distinct occurrences of the same data instances that provide conflicting information or are the data equivalent?

- Uniqueness:

1. Are data duplicated or do they have the unique feature for a field?
2. Do the data have duplicates, by mistake or do they have the characteristic of uniqueness for the dataset?

- Validity:

1. Does data correctly represent reality?
2. Are the data reliable?

- Integrity:

1. Is a dataset a measure of existence, validity, structure, content for the model?
2. Is the data correct?

- Timeliness:

1. Are data updated?
2. Do the data change with a high frequency?

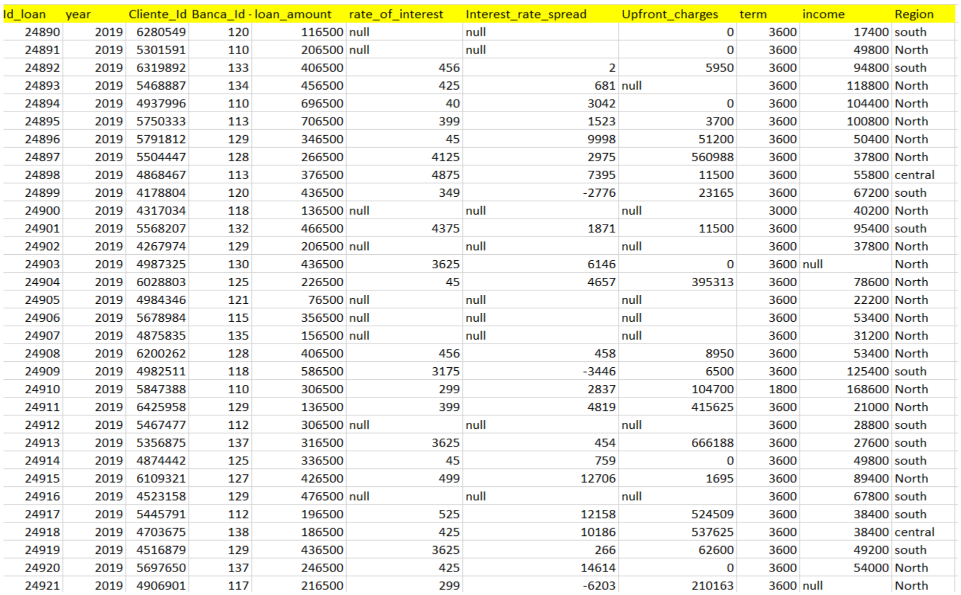
4.1 Validation results

The validation process is applied to four datasets. A set for loan and one for customer were chosen from Kaggle to be compared with two datasets created by us. The Kaggle one can be considered the benchmark dataset.

Specifically, a loan dataset from Kaggle [8] is evaluated as the best one for quality. A dataset preview is presented in Figure 2. The process of validation is divided in the following steps:

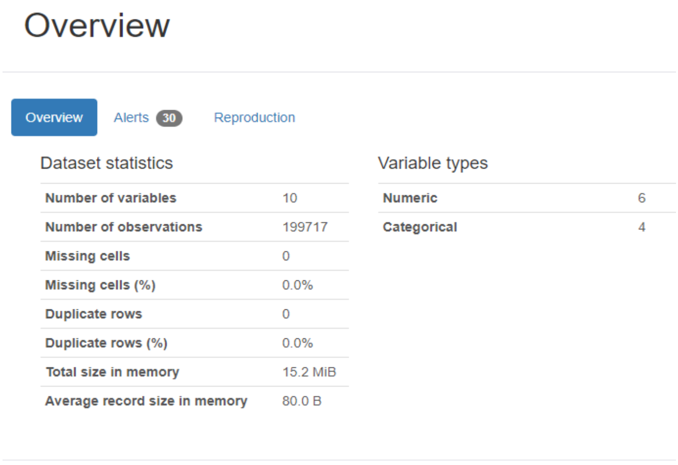
- Step 1: Pandas Profiling was applied to the dataset: in Figure 3 an overview of the Python analysis on this dataset is explained.
- Step 2: after this analysis, metrics were calculated. In Table 4 there is a presentation of results for each metric that contributes to score DQI questions.
- Step 3: eventually DQI can be scored, based on the metrics.

[8] https://www.kaggle.com/datasets/yasserh/loan-default-dataset



Id_loan	year	Cliente_Id	Banca_Id -	loan_amount	rate_of_interest	Interest_rate_spread	Upfront_charges	term	income	Region
24890	2019	6280549	120	116500	null	null	0	3600	17400	south
24891	2019	5301591	110	206500	null	null	0	3600	49800	North
24892	2019	6319892	133	406500	456	2	5950	3600	94800	south
24893	2019	5468887	134	456500	425	681	null	3600	118800	North
24894	2019	4937996	110	696500	40	3042	0	3600	104400	North
24895	2019	5750333	113	706500	399	1523	3700	3600	100800	North
24896	2019	5791812	129	346500	45	9998	51200	3600	50400	North
24897	2019	5504447	128	266500	4125	2975	560988	3600	37800	North
24898	2019	4868467	113	376500	4875	7395	11500	3600	55800	central
24899	2019	4178804	120	436500	349	-2776	23165	3600	67200	south
24900	2019	4317034	118	136500	null	null	null	3000	40200	North
24901	2019	5568207	132	466500	4375	1871	11500	3600	95400	south
24902	2019	4267974	129	206500	null	null	null	3600	37800	North
24903	2019	4987325	130	436500	3625	6146	0	3600	null	North
24904	2019	6028803	125	226500	45	4657	395313	3600	78600	North
24905	2019	4984346	121	76500	null	null	null	3600	22200	North
24906	2019	5678984	115	356500	null	null	null	3600	53400	North
24907	2019	4875835	135	156500	null	null	null	3600	31200	North
24908	2019	6200262	128	406500	456	458	8950	3600	53400	North
24909	2019	4982511	118	586500	3175	-3446	6500	3600	125400	south
24910	2019	5847388	110	306500	299	2837	104700	1800	168600	North
24911	2019	6425958	129	136500	399	4819	415625	3600	21000	North
24912	2019	5467477	112	306500	null	null	null	3600	28800	south
24913	2019	5356875	137	316500	3625	454	666188	3600	27600	south
24914	2019	4874442	125	336500	45	759	0	3600	49800	south
24915	2019	6109321	127	426500	499	12706	1695	3600	89400	North
24916	2019	4523158	129	476500	null	null	null	3600	67800	south
24917	2019	5445791	112	196500	525	12158	524509	3600	38400	south
24918	2019	4703675	138	186500	425	10186	537625	3600	38400	central
24919	2019	4516879	129	436500	3625	266	62600	3600	49200	south
24920	2019	5697650	137	246500	425	14614	0	3600	54000	North
24921	2019	4906901	117	216500	299	-6203	210163	3600	null	North

Figure 2: loan dataset overview



Overview

Dataset statistics

Number of variables	10
Number of observations	199717
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	15.2 MiB
Average record size in memory	80.0 B

Variable types

Numeric	6
Categorical	4

Figure 3: Dataset statistics (Kaggle dataset)

Table 3

Metrics results (Kaggle dataset)

Loan dataset 1	Metric	Question	Score (1-5)	Percentage
	Accuracy			**90%**
		1.	4	80%
		2.	5	100%
	Completeness			**90%**
		1.	5	100%
		2.	4	80%
	Consistency			**90%**
		1.	5	100%

		2.	4	80%
	Uniqueness			**100%**
		1.	5	100%
		2.	5	100%
	Validity			**80%**
		1.	4	80%
		2.	4	80%
	Integrity			**70%**
		1.	3	60%
		2.	4	80%
	Timeliness			**80%**
		1.	4	80%
		2.	4	80%

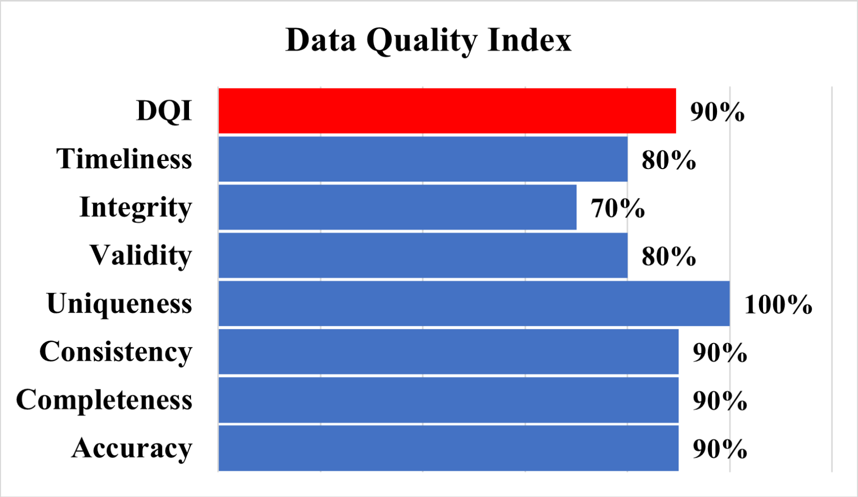


Figure 4: Data Quality Index (benchmark dataset)

Table 4
Metrics results (Our dataset)

Loan dataset 2	Metrics	Question	Score (1-5)	Percentage
	Accuracy			**60%**
		1.	3	60%
		2.	3	60%
	Completeness			**90%**
		1.	5	100%
		2.	4	80%
	Consistency			**80%**
		1.	4	80%
		2.	4	80%
	Uniqueness			**100%**
		1.	5	100%
		2.	5	100%

	Validity			**60%**
		1.	3	60%
		2.	3	60%
	Integrity			**70%**
		1.	3	60%
		2.	4	80%
	Timeliness			**50%**
		1.	3	60%
		2.	2	40%

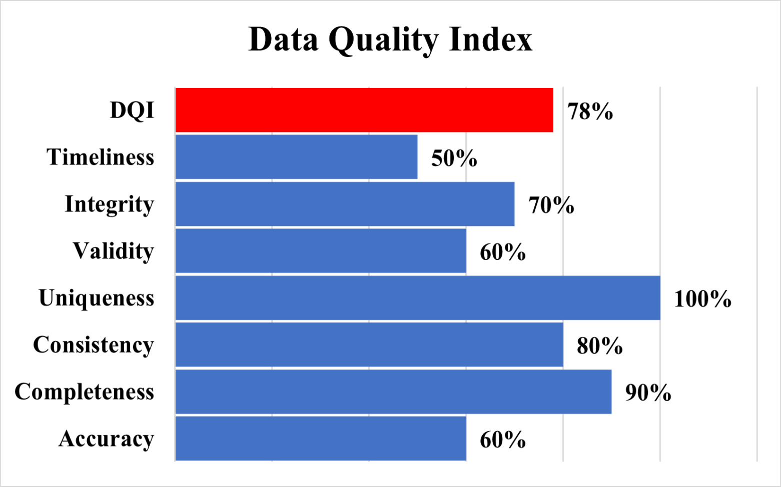


Figure 4: Data Quality Index(Our dataset)

From the analysis of the Data Quality Index, the best dataset is the loan one (91%), from Kaggle. Metrics are very positive because the dataset has the greatest number of observations; therefore, it is more complete and valid; it has got zero null and duplicate values that guarantee its uniqueness.

The second dataset, the customer one from Kaggle[9], scored a DQI of 85%, lower than the first, for the presence of about 6% of null values and a lower number of total values. The first two datasets present a better Timeliness because they are public and as such more current and updated than the others.

Finally, these results can be compared with two datasets (loan and customer) from our realization. As we expected, our datasets are smaller and less updated than the other ones, but they are more correct. They have a DQI of 78% and 80%, a high quality, slightly lower than the previous ones because the datasets are characterized by a limited number of observations (1000) and therefore have a lower level of completeness. Datasets are valid, with no null or duplicate values, ensuring greater uniqueness. IDEA gives us several opportunities because it minimizes data cleaning and normalization problems, but the DQI comparison done shows our model limits too.

As we can see the reason why our loan dataset has got a lower quality is due to accuracy, about 60%. Moreover, validity results 60% as the amount of data that makes the dataset representative of reality is 1000 records. Timeliness for 50%, due to a slow data update frequency.

Our model could be improved, by generating a higher number of records which should be updated every year. What we want to explain is that all the operations were made by hand, and it causes model limits. Our aim is to make the model more efficient thanks to Entity resolution tools which can automate manual operations by reducing manual errors and optimizing the working time.

[9] https://www.kaggle.com/

5. Conclusion

In this paper, an approach to the realization of realistic data sets for Artificial Intelligence was presented (IDEA). The approach was aimed at solving business needs and user requests. The main feature of our approach is an analysis to create a model applicable to the background.

By examining literature, we achieved that the main method applied was SEMMA. We aimed to make IDEA an extension of SEMMA method. IDEA can be useful for every reality and the use case chosen is the banking creditworthiness. The paper explained the use case and its validation. Specifically, this model can be used by every bank or financial institution to forecast the solvency of a customer portfolio. As explained our purpose was developing a user centered approach to understand business requests. A business needs an efficient strategy that can be improved thanks to IDEA because our approach gives a representation of a business reality from two points of view: customers and businesses.

6. References

[1] Alam, Talha Mahboob, Kamran Shaukat, Ibrahim A. Hameed, Suhuai Luo, Muhammad Umer Sarwar, Shakir Shabbir, Jiaming Li, and Matloob Khushi. “An Investigation of Credit Card Default Prediction in the Imbalanced Datasets.” *IEEE Access* 8 (2020): 201173–98. https://doi.org/10.1109/ACCESS.2020.3033784.

[2] Azevedo, Ana, and Manuel Filipe Santos. “KDD, Semma and CRISP-DM: A Parallel Overview.” *MCCSIS’08 - IADIS Multi Conference on Computer Science and Information Systems; Proceedings of Informatics 2008 and Data Mining 2008*, no. January 2008 (2008): 182–85.

[3] Madaan, Mehul, Aniket Kumar, Chirag Keshri, Rachna Jain, and Preeti Nagrath. “Loan Default Prediction Using Decision Trees and Random Forest: A Comparative Study.” *IOP Conference Series: Materials Science and Engineering* 1022, no. 1 (2021). https://doi.org/10.1088/1757-899X/1022/1/012042.

[4] Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. “An Approach for Prediction of Loan Approval Using Machine Learning Algorithm.” *Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020*, no. Icesc (2020): 490–94. https://doi.org/10.1109/ICESC48915.2020.9155614.

[5] Sivasree M S, and Rekha Sunny T. “Loan Credibility Prediction System Based on Decision Tree Algorithm.” *International Journal of Engineering Research And* V4, no. 09 (2015): 825–30. https://doi.org/10.17577/ijertv4is090708.

[6] Tariq, Hafiz Ilyas, Asim Sohail, Uzair Aslam, and Nowshath Kadhar Batcha. “Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess (Semma).” *Journal of Computational and Theoretical Nanoscience* 16, no. 8 (2019): 3489–3503. https://doi.org/10.1166/jctn.2019.8313.

[7] Tejaswini, J, T Mohana Kavya, R Devi, Naga Ramya, P Sai Triveni, and Venkata Rao Maddumala. “ACCURATE LOAN APPROVAL PREDICTION BASED ON MACHINE LEARNING APPROACH” 11 (2020). www.jespublication.com.

[8] Turiel, J. D., and T. Aste. “Peer-to-Peer Loan Acceptance and Default Prediction with Artificial Intelligence: P2P Default Prediction with AI.” *Royal Society Open Science* 7, no. 6 (2020). https://doi.org/10.1098/rsos.191649rsos191649.

[9] Zhu, Lin. “ScienceDirect A Study Study on Predicting Loan Default Based on the Random Forest Algorithm.” *Procedia Computer Science* 162, no. Itqm 2019 (2020): 503–13. https://doi.org/10.1016/j.procs.2019.12.017.

[10] Sidi, Fatimah, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim, and Aida Mustapha. “Data Quality: A Survey of Data Quality Dimensions.” *Proceedings - 2012 International Conference on Information Retrieval and Knowledge Management, CAMP’12*, no. August (2012): 300–304. https://doi.org/10.1109/InfRKM.2012.6204995.