Assignment 1

Anonymization of Datasets with Privacy, Utility and Risk Analysis

Gonçalo Melo
Department of Computer Science
Faculty of Sciences
University of Porto
Porto, Portugal
up202308365@up.pt

Martim Ribeiro
Department of Computer Science
Faculty of Sciences
University of Porto
Porto, Portugal
up202308304@up.pt

***Abstract*—This paper presents a systematic approach to anonymizing a dataset sourced from Kaggle, tailored for two-wheeler loan profiles in India. Beginning with the selection and importing phases, we meticulously curated the dataset to ensure relevance and integrity. Subsequently, attributes were classified and characterized into sensitive, quasi-identifying, and insensitive categories, followed by the definition of coding models and hierarchies. Finally, we evaluated and compared multiple privacy models to strike a balance between data utility and privacy risk. Our analysis sheds light on the challenges and trade-offs inherent in anonymization processes, offering valuable insights for practitioners and researchers navigating similar endeavors.**

I. Introduction

II. Selection, Importing and Goal of the Dataset

A. Selection of the Dataset

In our pursuit of a suitable dataset for anonymization purposes, we meticulously sought out data with specific criteria in mind. Firstly, our primary consideration was the dataset's relevance in the context of anonymization and its aptness for the application of anonymization techniques. Understanding the critical importance of safeguarding personal data, we aimed to procure a dataset that not only resonated with the principles of privacy protection but also provided a practical ground for implementing anonymization methodologies effectively.

Additionally, we aimed for the dataset to be anchored in the domain of banking or finance, recognizing the inherent sensitivity and richness of financial data for anonymization endeavors.

Our search journey led us through suggested sources renowned for their diverse collection of datasets, with Kaggle emerging as our ultimate destination. Kaggle's platform, known for its vast repository of high-quality datasets and active community engagement, offered us a fertile ground for exploration. Within Kaggle's extensive collection, we unearthed a dataset that aligned with our requisites—a dataset detailing credit loans specifically tailored for two-wheelers (https://www.kaggle.com/datasets/yashkmd/credit-profile-two-wheeler-loan-dataset).

B. Importing the Dataset

Upon importing the dataset into ARX, it was observed that the dataset was already relatively clean and did not require extensive sanitization procedures. This allowed us to streamline the import process and focus on other aspects of the anonymization workflow.

However, one notable observation was the considerable size of the dataset, which could potentially make anonymization operations computationally expensive and time-consuming. To address this challenge, a decision was made to work with a representative sample of the dataset, approximately 20% of the total records. By working with a smaller subset of the data, we could expedite the testing and evaluation of anonymization techniques while still maintaining the integrity and representativeness of the original dataset.

C. Goal of the Dataset

The goal of releasing the anonymized dataset is to provide a comprehensive overview of potential loan applicants' profiles, specifically tailored for the Indian demographic, without disclosing information about individuals that could lead to re-identification attacks, thereby safeguarding their privacy.

The dataset encompasses a diverse range of features, including basic demographics and financial details, which are crucial for evaluating an individual's creditworthiness.

III. Characterization of the Dataset and Coding Models

A. Attribute Classification

Upon importing the dataset into ARX, it is necessary to classify the attributes accordingly. The process of categorizing attributes into `insensitive`, `sensitive`, `quasi-identifying (QID)` or `identifying` is often subjective and depends on the goal of releasing the dataset and the privacy requirements. In this subsection, we present the decisions made regarding the classification of attributes and justify the choices behind them. Note that, given that this dataset was retrieved from a publicly available source (Kaggle), it doesn't contain any explicit identifying attributes that could immediately compromise privacy requirements, as is often the case with datasets from such sources. Furthermore, the decision between classifying an attribute as a QID or not was supported by the values of distinction and separation provided in ARX.

- **Age:** `QID` - Age is often classified as a quasi-identifying attribute (QID) due to its potential to indirectly reveal an individual's identity when combined with other attributes. While age alone may not uniquely identify a person, it can significantly narrow down the pool of potential individuals when combined with other quasi-identifiers such as ZIP code, gender, or occupation. Indeed, as we'll see, some of these other attributes which are commonly classified as QIDs are also present in the dataset.
- **Gender:** `QID` - Just like age, when combined with other demographic information, it can contribute to re-identification risks. Additionally, societal norms and distributions of gender across populations can further reduce anonymity, making it a relevant factor to consider in privacy protection measures.
- **Income:** `Sensitive` - Income is considered a highly sensitive attribute as it directly reflects an individual's earning capacity, socioeconomic status, and financial privacy. Income disclosure may carry risks of discrimination, stigma, and privacy breaches.
- **Credit Score** and **Profile Score:** `Sensitive` - While credit score and profile score may not directly reveal an individual's income, they still convey sensitive information about an individual's financial behavior, creditworthiness, and overall financial health. These scores are often used by lenders and financial institutions to assess the risk of lending money to individuals. Therefore, they have implications for an individual's access to financial services, loan terms, and interest rates. Additionally, credit score and profile score can indirectly reflect an individual's financial stability and ability to manage debt, which may impact their socioeconomic status and overall well-being.
- **Credit History Length:** `QID` - Credit history length provides insights into an individual's financial behavior over time.
- **Number of Existing Loans:** `QID` - The number of existing loans of an individual can also be combined with other data for re-identification purposes.
- **Loan Amount**: and **Loan Tenure:** `Insensitive` - Loan values and loan tenures for two-wheeler loans can be classified as insensitive attributes due to their potential variance even within the same individual's borrowing history. Individuals may have taken out loans for two-wheelers with significantly different prices and repayment periods over time. For example, an individual might have borrowed money for two-wheelers with very distinct prices, such as a small scooter for daily commuting and a high-end motorcycle for recreational purposes. This variability within an individual's borrowing history makes it challenging to rely on loan values and tenures for re-identification purposes. Even within the specific context of two-wheeler loans, the range of possible loan values and tenures reflects diverse financial circumstances, preferences, and needs among borrowers. Consequently, the substantial variance in loan values and tenures within an individual's borrowing history mitigates the risk of these attributes being sensitive identifiers, as they do not consistently or uniquely represent an individual's identity across different loan transactions for two-wheelers.
- **Existing Customer:** `QID` - "Existing Customer" could be considered quasi-identifying, especially if combined with other attributes. For example, knowing that someone is an existing customer of a particular bank or retailer could narrow down their identity when combined with other demographic information.
- **State:** `QID` - While state alone may not uniquely identify an individual, it significantly reduces the pool of potential candidates, particularly in countries with large populations like India. For instance, in India, knowing the state can often narrow down the search to a relatively small group of individuals.
- **City:** `QID` - Similarly, city information further narrows down the pool of potential candidates. Cities in India are usually more densely populated, so knowing the city in addition to the state can substantially reduce anonymity. Additionally, certain cities may have unique demographic characteristics or be associated with specific industries or cultural attributes, making them more identifiable.
- **LTV Ratio:** `Sensitive` - The Loan-to-Value (LTV) ratio is considered a sensitive attribute as it not only reveals information about the loan amount but also indirectly discloses the value of the two-wheeler being financed.
- **Employment Profile:** `QID` - For this specific context, we decided to classify employment profile as a quasi-identifier. For example, knowing that an individual is self-employed in a specific industry, such as technology or healthcare, along with other demographic details like age and location, could significantly reduce the number of possible matches and increase the risk of re-identification.
- **Occupation:** `QID` - Knowing the occupation of an individual can significantly contribute to narrowing down the pool of potential candidates, especially in smaller or more specialized fields. Occupations often have distinct characteristics, skill sets, and income levels associated with them, making individuals within the same occupation group more easily distinguishable.

B. Attribute Distribution

In the analysis of the dataset, it was observed that all values follow a reasonable distribution and with no outliers, indicating consistency and reliability in the data. This ensures that the dataset accurately represents the underlying characteristics being studied, enhancing the validity of subsequent analyses.

Regarding the sensitive attribute, the Loan-to-Value (LTV) ratio, it is imperative to assess whether its high decimal precision necessitates approximation for anonymization models. Since LTV ratio values exhibit excessive decimal precision, it may increase the risk of re-identification when used in anonymization models. In such cases, rounding or approximation techniques may be employed to reduce the precision

Fig. 1: Privacy Risks of the Dataset in the Original Form

of the LTV ratio values while preserving the overall trends and patterns in the data.

C. Privacy Risks

The original form of the dataset presents a high privacy risk for the several attacker models (prosecutor, journalist and marketer), as highlighted by the risk analysis conducted using the ARX tool. The risk analysis report generated by ARX, presented in Fig. 1 highlights the critical need for thorough anonymization processes to safeguard individual privacy while enabling meaningful analysis of the dataset.

D. Coding Models

In this subsection, we explore the hierarchies established through generalization techniques and attribute weights configured within ARX. These methods play a vital role in enhancing privacy protection while attempting to preserve data utility.

- **Age:** Intervals incremented by a value of 10;
- **Gender:** Generalized to a common value through a fully generalized hierarchy, encompassing categories of male, female, and other;
- **Credit History Length:** Intervals increment by a value of 50 and top coding for values $>= 600$;
- **Number of Existing Loans:** Intervals increment by a value of 2 and top coding for values $>= 10$;
- **Existing Customer**: Generalized to a common value through a fully generalized hierarchy, encompassing the values ”No” and ”Yes”;
- **State** and **City:** Hierarchies encompassing different levels of generalization in the following order: *State -> Region (North, Central, South) -> Country (India)*;
- **Employment Profile** and **Occupation:** Hierarchies encompassing different levels of generalization in the following order: (Unemployed, Employed) -> Wildcard (*)

Level-0	Level-1	Level-2	Level-3	Level-4
18	[10, 20[	[0, 20[	[0, 40[	[0, 80[
19	[10, 20[	[0, 20[	[0, 40[	[0, 80[
20	[20, 30[	[20, 40[	[0, 40[	[0, 80[
21	[20, 30[	[20, 40[	[0, 40[	[0, 80[
22	[20, 30[	[20, 40[	[0, 40[	[0, 80[
23	[20, 30[	[20, 40[	[0, 40[	[0, 80[
24	[20, 30[	[20, 40[	[0, 40[	[0, 80[
25	[20, 30[	[20, 40[	[0, 40[	[0, 80[
26	[20, 30[	[20, 40[	[0, 40[	[0, 80[
27	[20, 30[	[20, 40[	[0, 40[	[0, 80[
28	[20, 30[	[20, 40[	[0, 40[	[0, 80[
29	[20, 30[	[20, 40[	[0, 40[	[0, 80[
30	[30, 40[	[20, 40[	[0, 40[	[0, 80[
31	[30, 40[	[20, 40[	[0, 40[	[0, 80[
32	[30, 40[	[20, 40[	[0, 40[	[0, 80[
33	[30, 40[	[20, 40[	[0, 40[	[0, 80[
34	[30, 40[	[20, 40[	[0, 40[	[0, 80[
35	[30, 40[	[20, 40[	[0, 40[	[0, 80[
36	[30, 40[	[20, 40[	[0, 40[	[0, 80[
37	[30, 40[	[20, 40[	[0, 40[	[0, 80[
38	[30, 40[	[20, 40[	[0, 40[	[0, 80[
39	[30, 40[	[20, 40[	[0, 40[	[0, 80[

Fig. 2: Age Hierarchies

Level-0	Level-1
Female	{Female, Male, Other}
Male	{Female, Male, Other}
Other	{Female, Male, Other}

Fig. 3: Gender Hierarchies

Figures 2 through 8 showcase the hierarchies chosen for each quasi-identifier.

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5	Level-6
6	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
7	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
8	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
9	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
10	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
11	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
12	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
13	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
14	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
15	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
16	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
17	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
18	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
19	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
20	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
21	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
22	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
23	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
24	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
25	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
26	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*
27	[0, 50[	[0, 100[	[0, 200[	[0, 400[	[0, 600[	*

Fig. 4: Credit History Length Hierarchies

Level-0	Level-1	Level-2	Level-3	Level-4	Level-5
0	[0, 2[	[0, 4[	[0, 8[	[0, 10[	*
1	[0, 2[	[0, 4[	[0, 8[	[0, 10[	*
2	[2, 4[	[0, 4[	[0, 8[	[0, 10[	*
3	[2, 4[	[0, 4[	[0, 8[	[0, 10[	*
4	[4, 6[	[4, 8[	[0, 8[	[0, 10[	*
5	[4, 6[	[4, 8[	[0, 8[	[0, 10[	*
6	[6, 8[	[4, 8[	[0, 8[	[0, 10[	*
7	[6, 8[	[4, 8[	[0, 8[	[0, 10[	*
8	[8, 10[	[8, 10[	[8, 10[	[0, 10[	*
9	[8, 10[	[8, 10[	[8, 10[	[0, 10[	*
10	>=10	>=10	>=10	>=10	*

Fig. 5: Number of Existing Loans Hierarchies

Level-0	Level-1
No	{No, Yes}
Yes	{No, Yes}

Fig. 6: Existing Customer Hierarchies

Level-0	Level-1	Level-2
Rajasthan	North	India
Delhi	North	India
Uttar Pradesh	North	India
West Bengal	North	India
Gujarat	Central	India
Maharashtra	Central	India
Karnataka	South	India
Kerala	South	India
Tamil Nadu	South	India
Telangana	South	India

(a) State Hierarchies

Level-0	Level-1	Level-2	Level-3
New Delhi	Delhi	North	India
Kanpur	Uttar P...	North	India
Lucknow	Uttar P...	North	India
Bishanpura	Uttar P...	North	India
Udaipur	Rajast...	North	India
Jaipur	Rajast...	North	India
Dhulagori	West B...	North	India
Kolkata	West B...	North	India
Ahmedabad	Gujarat	Central	India
Surat	Gujarat	Central	India
Mumbai	Mahar...	Central	India
Pune	Mahar...	Central	India
Manjari	Mahar...	Central	India
Nagpur	Mahar...	Central	India
Channarayap...	Karnat...	South	India
Bengaluru	Karnat...	South	India
Mysuru	Karnat...	South	India
Kochi	Kerala	South	India
Thiruvanant...	Kerala	South	India
Chennai	Tamil ...	South	India
Nellikuppam	Tamil ...	South	India
Coimbatore	Tamil ...	South	India

(b) City Hierarchies

Fig. 7: State and City Hierarchies

Level-0	Level-1	Level-2
Self-Employed	Employed	*
Freelancer	Employed	*
Salaried	Employed	*
Student	Unemployed	*
Unemployed	Unemployed	*

(a) Employment Profile Hierarchies

Level-0	Level-1	Level-2
	Unemployed	*
Student	Unemployed	*
Banker	Employed	*
Business Ow...	Employed	*
Civil Servant	Employed	*
Contractor	Employed	*
Doctor	Employed	*
Farmer	Employed	*
Graphic Desi...	Employed	*
Independent...	Employed	*
Photographer	Employed	*
Shopkeeper	Employed	*
Software En...	Employed	*
Teacher	Employed	*
Writer	Employed	*

(b) Occupation Hierarchies

Fig. 8: Employment Profile and Occupation Hierarchies

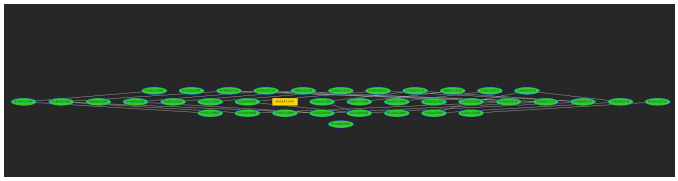

Fig. 9: Results Lattice for the First Model

Regarding attribute weights, they were kept with the default value of 0.5 at the current stage. However, re-definitions of these values will be considered while facing iterations over the privacy models.

IV. Privacy Models: Utility, Privacy and Risk Assessment

In this section, we delve into the critical aspects surrounding the selection and evaluation of privacy models within the context of anonymization. We commence by discussing the process of choosing privacy models, where considerations such as dataset characteristics, privacy requirements, and utility preservation are meticulously examined. Following this, a detailed analysis is conducted to evaluate the performance of each selected privacy model. This examination entails assessing their effectiveness in safeguarding individual privacy while balancing data utility.

Moreover, a detailed utility and risk analysis is undertaken to scrutinize the impact of privacy-preserving techniques on the overall utility of the anonymized dataset and the mitigated risks of re-identification attacks. By meticulously examining these factors, we endeavor to provide insights that inform strategic decision-making in the pursuit of achieving optimal data privacy and utility.

A. First Model: 3-Anonymity and 2-Diversity

Our first privacy model employs a combination of k-anonymity and l-diversity techniques.

Specifically, for quasi-identifiers, we apply k-anonymity with a parameter value of **k=3**. This ensures that each group of records in the dataset is indistinguishable from at least two other records with respect to the specified quasi-identifiers.

Additionally, for sensitive attributes, we implement l-diversity with a parameter of **l=2**, which guarantees that each equivalence class contains at least two distinct values of the sensitive attribute. This ensures that sensitive attribute values are sufficiently diverse within each group of records, preventing potential attribute disclosure and enhancing privacy protection.

The results provided by ARX are visable in Fig. 9

The lattice shows us that the optimal hierarchies level chosen were:

- **Age:** Level 4
- **Gender:** Level 0
- **Credit History Length:** Level 6
- **Number of Existing Loans:** Level 5
- **Existing Customer:** Level 0

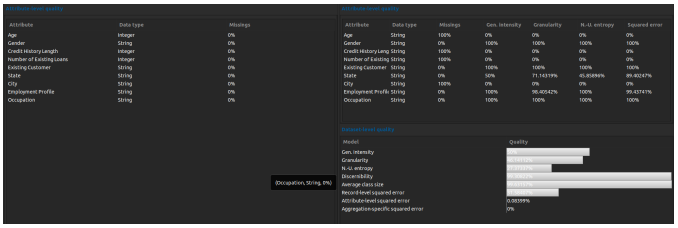

Fig. 10: Quality Analysis of the Anonymized Dataset

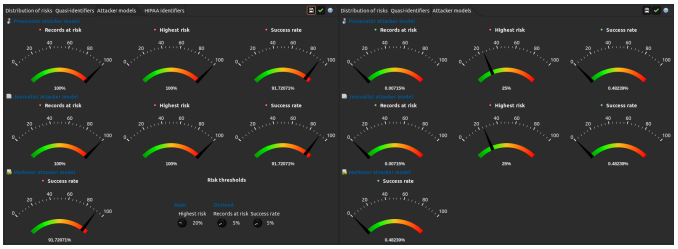

Fig. 11: Risk Analysis of the Anonymized Dataset

- **State:** Level 1
- **City:** Level 3
- **Employment Profile:** Level 0
- **Occupation:** Level 0

We can interpret these results by understanding that the critical attributes that were chosen to be highly anonymized were **Age**, **Credit History Length**, **Number of existing Loans** and **City**.

The selected parameters of k-anonymity l-diversity represent the lowest levels of anonymization specified. While prioritizing utility, these parameters also ensure an acceptable level of privacy protection. By striking a balance between data utility and privacy, the chosen parameters achieve meaningful anonymization without excessive distortion of the dataset as demonstrated in Figures 10 and 11. Keep in mind that the algorithm used to evaluate quality over the anonymized dataset was Logistic Regression. This approach underscores a pragmatic trade-off between utility and privacy, ensuring that the anonymized dataset might remain useful for analysis.

Fig. 11 specifically showcases the extremely low success rate for different attacker models (about 0.5% success rate).

An iterative process of redefining attribute weights was performed in order to attempt achieving levels of less generalization in some the attributes such as **Age** and **Number of Existing Loans** which were anonymized with the most general level of the hierarchies. The results, however, did not change and appeared to have no impact in the choices of the privacy model, so it was decided to keep an even attribute weight distribution.

B. Second Model: 3-Anonimity and 0.9-Closeness

The second privacy model combines the principles of 3-anonymity and 0.9-closeness to ensure robust privacy protection while preserving data utility.

We employ 0.9-closeness on each sensitive attribute to ensure that the distribution of sensitive attribute values within each equivalence class closely matches the overall distribution in the dataset. With a closeness threshold of **0.9**, we aim to minimize the discrepancy between local and global distributions, thereby preserving the statistical properties of the sensitive attribute while protecting individual privacy.

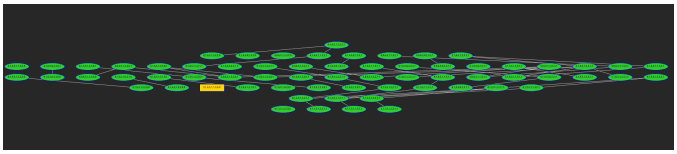

Fig. 12: Results Lattice for the Second Model

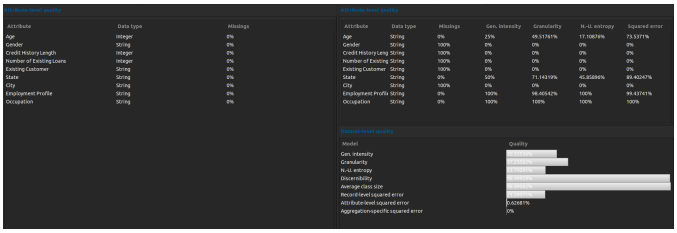

Fig. 13: Quality Analysis of the Anonymized Dataset

Fig. 12 represents the results in the lattice form. Fig. 13 showcases a slightly decrease on the quality preserved in comparison to the first model. Despite that, a trade-off between utility and privacy can be observed on Fig. 14, that demonstrates a decrease in the risk of re-identification regarding the different attacker models (about 0.16% success rate).

The lattice shows us that the optimal hierarchies level chosen were:

- **Age:** Level 3
- **Gender:** Level 1
- **Credit History Length:** Level 6
- **Number of Existing Loans:** Level 5
- **Existing Customer:** Level 1
- **State:** Level 1
- **City:** Level 3
- **Employment Profile:** Level 0
- **Occupation:** Level 0

Like in the previous model, we see a pattern in the fact that it was preferred to highly generalize attributes like **Age**, **Credit History Length**, **Number of Existing Loans** and **City**, in comparison to attributes like **Employment Profile** and **Occupation**, that were kept with more specific information. However, the choices for generalization were more profound, resulting in the loss of some utility in relation to the previous model. Still, we can infer that the utility of the dataset regarding the comprehensive analysis of the profile of a loan applicants remains possible, especially regarding their employment profile and the state where they live in India.

The decision to cap the parameter k in k-anonymity at a maximum value of 3 on each privacy model was driven by the satisfactory privacy levels already achieved and the desire to avoid compromising data utility. While higher values of

Fig. 14: Risk Analysis of the Anonymized Dataset

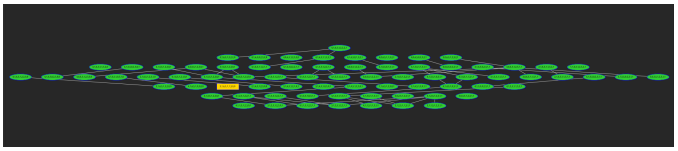

Fig. 15: Results Lattice of the Anonymized Dataset

k might offer increased privacy, they risk diminishing the dataset's granularity and analytical value. By maintaining k at 3, we struck a balance between privacy and utility, ensuring robust protection against re-identification while preserving meaningful data for analysis and decision-making purposes.

Similarly, it was also decided to keep the value of the parameter l as 2 in l-diversity and the value of threshold t as 0.9 for t-closeness, as those were the values that offered the best trade-off in terms of privacy and utility

C. Third Model: Combining Different Models

For a third model, we took into consideration that achieving 2-diversity for a highly decimal-precise sensitive attribute like the Loan-to-Value (LTV) ratio is relatively straightforward due to the abundance of distinct values resulting from its precision. So, we opted to change the First Model and change the algorithm of the sensitive attribute **LTV Ratio** to 0.9-closeness.

The results (Figures 15 throught 17) were, curiously, extremely similar as the ones from the Second Model where only 3-anonimity and 0.9-closeness was applied.

Additionaly, in assessing the success of anonymization within ARX, one crucial metric to consider is the reduction in values of distinction and separation compared to the original dataset.

By observing reductions in both values of distinction and separation, it can be inferred that the anonymization process has succeeded in obscuring individual identities and group

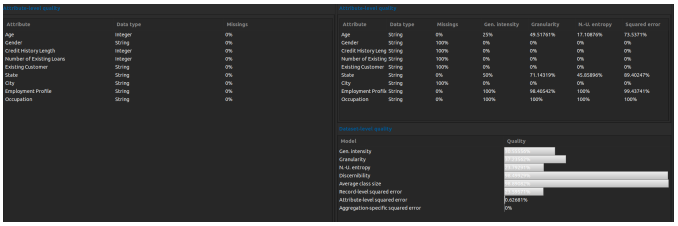

Fig. 16: Quality Analysis of the Anonymized Dataset

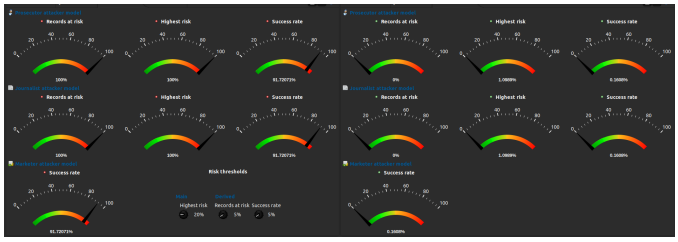

Fig. 17: Risk Analysis of the Anonymized Dataset

characteristics, thus enhancing privacy while preserving data utility. This serves as a key indicator of the effectiveness of the anonymization techniques employed within ARX.

V. Conclusion

In conclusion, our exploration into anonymization models revealed that there is no singular best approach for all scenarios. Instead, we encountered a trade-off between privacy and utility, with each model offering varying degrees of protection and analytical value. While some models prioritized privacy, others leaned towards preserving utility. Ultimately, the choice of model depends on the specific requirements of the dataset and the intended use case.

In our case, we found that satisfactory privacy levels could be achieved even with models that offered greater utility, such as the First Model. The First Model was able to maintain better analytical value, without compromising privacy and allowing easy re-identification- Given this outcome, we opt for models that strike a balance between privacy and utility, ensuring that the anonymized dataset remains valuable for analysis while adequately safeguarding individual privacy. This decision underscores the importance of tailoring anonymization strategies to the unique characteristics and goals of each dataset, thereby maximizing its usefulness while upholding privacy principles.

REFERENCES

[1] ARX Documentation https://arx.deidentifier.org/
[2] Kaggle Webiste https://kaggle.com/
[3] Kaggle Chosen Dataset https://www.kaggle.com/datasets/yashkmd/credit-profile-two-wheeler-loan-dataset