Class-wise confidence for debt prediction in real estate management: discussion and lessons learned from an application

Soundouss Messoudi soundouss.messoudi@hds.utc.fr
Sébastien Destercke sebastien.destercke@hds.utc.fr
Sylvain Rousseau sylvain.rousseau@hds.utc.fr
HEUDIASYC - UMR CNRS 7253, Université de Technologie de Compiègne, 57 avenue de Landshut, 60203 COMPIEGNE CEDEX - FRANCE

Editor: Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin and Khuong An Nguyen

Abstract

The prediction of tenants likely to fall into a debt situation is a key issue for social property owners in real estate. It is even more important for them to limit the number of people falsely predicted to be in debt to avoid incurring unnecessary costs (in time and money), for instance by sending agents to prevent the debt. In this paper, we adapt mondrian conformal prediction to control the error rate of this class, while keeping a level of confidence chosen by the social property owner, or more generally by the user. We also test this small adaptation with different splitting strategies and discuss the obtained results, those later showing promising results, in the sense that they show that our approach can work, as well as pointing out and discussing difficulties, in the sense that conformal prediction fails on some settings of particular interest to the end-user.

Keywords: Inductive conformal prediction, mondrian conformal prediction, class-wise confidence, debt classification, real estate.

1. Introduction

Social housing is housing intended for people with modest incomes who would have difficulty finding housing on the private market. The social lease is granted under conditions of income or family composition. Even if these rents are smaller than the average rents in the geographic sector, the social housing remains a victim of unpaid rent. The generating facts that explain these unpaid situations originate from predictable reasons (job insecurity, multiple consumer loans, etc.) or unpredictable (tight budget, health problems, change in family situations such as a birth or a divorce...). In addition, the number of households in debt is likely to increase with the health crisis due to the COVID-19 pandemic (Manville et al. (2020)). In each case, social property owners must study the particular situation of the household in difficulty of payment and hire social counsellors and litigation managers to help them resolve their problems before reaching an eviction phase.

Thus, it is essential for social property owners to anticipate tenants that are likely to fall into debt, and more importantly, limit the number of tenants that are misclassified as in debt in order to avoid paying unnecessary costs, or wasting some social agent time that could otherwise have been used to the benefit of tenants really in need. On the other side, while it is important to maintain a good overall accuracy, the accuracy on the debt class is not so important, as misclassifying a tenant as having no debt only delays the recovering procedure. To control all these factors, we propose a class-wise confidence approach based on Mondrian conformal prediction.

Mondrian conformal prediction is a variant of conformal prediction, a framework that provides a statistical guarantee by giving a set of classes in the classification case and a prediction interval in the case of regression. One of the desirable features in conformal predictors is validity, i.e. the error rate does not exceed a probability error chosen by the user. In the case of Mondrian conformal predictors, validity is verified within categories of the data set instead of the global data set. The principle of conformal prediction and its Mondrian form for classification in the inductive setting will be recalled in Section 2.

Our work uses Mondrian conformal classification to get class-wise confidence applied to debt prediction in real estate management. By class-wise confidence, we mean that we are not aiming at achieving the same accuracy for all classes, e.g., for cost-sensitive reasons. Our approach is described in Section 3. The experiments and their results are presented in Section 4, where we adapt a standard Mondrian non-conformity measure to control the error of the class that is most important for the social property owner, while guaranteeing a global confidence. While some of our results demonstrate the feasibility of our approach and the fact that it can work, some others point out significant difficulties related to the application to reach validity. We discuss those as well as our next step to correct them.

2. Mondrian conformal prediction

In the case of classification, conformal prediction is a framework that provides a statistical guarantee of the coverage of the true class by predicting a subset of all classes with a confidence level defined by the user. This framework was first developed for a transductive online setting (Gammerman et al. (1998); Vovk et al. (2005)) that needs to train the model each time a prediction is sought, and then was adapted to the inductive setting (Papadopoulos (2008)) by keeping a set of the training examples for calibration instead of using them all to train the underlying algorithm. These classical conformal prediction methods guarantee an overall confidence level. Another approach called Mondrian conformal prediction (Vovk et al. (2003)) was proposed to get a statistical guarantee on a subset of the data set based on a condition. This section presents this method and some related works in classification.

2.1. Inductive conformal prediction (ICP) for classification

Let $z_1 = (x_1, y_1), \dots, z_n = (x_n, y_n) \in \mathbf{Z}$ be successive pairs constituting the examples, with $x_i \in \mathbf{X}$ an object and $y_i \in \mathbf{Y} = \{C_1, \dots, C_p\}$ its class. Let $\mathbf{Z}$ be exchangeable (a weaker condition than i.i.d.). We can predict $y_{n+1} \in \mathbf{Y}$ for any new object $x_{n+1} \in \mathbf{X}$ by following the inductive conformal framework steps:

1. Split the original data set $\mathbf{Z}$ into a *proper training set* $\mathbf{Z}^{tr}$ with $|\mathbf{Z}^{tr}| = m$ and a *calibration set* $\mathbf{Z}^{cal}$ with $|\mathbf{Z}^{cal}| = n - m = q$.

2. Train a classification *underlying algorithm* $h : \mathbf{X} \to \mathbf{Y}$ on $\mathbf{Z}^{tr}$ to obtain the *non-conformity measure* $f(z)$. The standard non-conformity measure in a classification

setting when h is a probabilistic classifier is defined as follows:

$$f(z) = 1 - \hat{P}_h[y \mid x]. \tag{1}$$

3. Apply the non-conformity measure $f(z)$ to each example z_i of $\mathbf{Z}^{cal}$ to get the non-conformity scores $\alpha_1, \dots, \alpha_q$.

4. Choose a *significance level* $\epsilon \in (0, 1)$ to get a prediction set with a *confidence level* of $1 - \epsilon$.

5. For a new example x_{n+1}, compute a non-conformity score for each class $C_k \in \mathbf{Y}$:

$$\alpha_{n+1}^{C_k} = f((x_{n+1}, y = C_k)). \tag{2}$$

6. For each class $C_k \in \mathbf{Y}$, compute the p-value :

$$p_{n+1}^{C_k} = \frac{|\{i \in 1, \dots, q : \alpha_{n+1}^{C_k} \leq \alpha_i\}|}{q}. \tag{3}$$

7. Build the prediction set :

$$\Gamma^\epsilon = \{C_k \in \mathbf{Y} : p_{n+1}^{C_k} > \epsilon\}. \tag{4}$$

The prediction set can be a singleton when the predictor is sure, a set with more than one class in case of ambiguity and an empty set $\emptyset$ when the model does not know or did not see a similar example during training. The two desirable properties in conformal predictors are (a) *validity*, i.e. the error rate does not exceed ϵ for each chosen confidence level $1 - \epsilon$, and (b) *efficiency*, meaning prediction sets are as small as possible.

2.2. Mondrian conformal prediction (MCP)

As mentioned above, the validity feature desired in a conformal classifier guarantees that the overall confidence level is maintained on all the data set by choosing an error rate ϵ that should not be exceeded. There is however no guarantee on a subset of the data set, or on specific categories of the data set. Mondrian conformal prediction provides this subset validity based on categories such as class-conditional or attribute-conditional categories (Vovk et al. (2003)). In this case, each category has its own individual guarantee based on the chosen individual significance level. For example, a class-conditional conformal classifier will give a validity on each class, whereas an attribute-conditional conformal classifier will guarantee the error rate for each category of the chosen attribute used for dividing the data set. We focus on the class-conditional form of Mondrian conformal predictors as it is what we use in our paper.

The difference between an inductive conformal classifier and a class-conditional conformal classifier is in the computation of the p-value (3), in which, instead of taking all non-conformity scores α_i in the calibration set, we only consider those related to the examples belonging to the same class we are hypothetically testing for the object x_{n+1}. The p-value becomes:

$$p_{n+1}^{C_k} = \frac{|\{i \in 1, \dots, q : y_i = C_k, \alpha_{n+1}^{C_k} \leq \alpha_i\}|}{|\{i \in 1, \dots, q : y_i = C_k\}|}. \tag{5}$$

Class-conditional MCP is mostly used when data is imbalanced, in order to maintain the same error rate even for the minority class.

Mondrian conformal prediction has been applied to a variety of real-world problems. In medicine, Yang et al. (2009) apply MCP for a cost-sensitive learning of medical diagnosis for Chronic Gastritis and Thyroid diseases, Candès et al. (2021) adapt an attribute-conditional conformal classifier for COVID-19 survival prediction, and Devetyarov et al. (2012) use class-conditional MCP to early diagnose cancer. In pharmaceutical research, Toccaceli and Gammerman (2019) combine p-values from different Mondrian conformal predictors to discover new drugs. Zhang et al. (2020) use MCP to reject unreliable answers for automatic Product Question Answering (PQA) in e-commerce applications. Our work adds another application to these many existing ones in the real estate sector.

3. Class-wise confidence: our approach

This section presents our approach for a class-wise confidence using Mondrian conformal prediction, in the case of a binary problem.

Our primary goal in this paper and the associated application is to control the error rate of a given class, while preserving a relatively low global error rate. A specificity of this case is that the error rate of the remaining class is of marginal importance, and that is to some extent what really matters for this second class is the efficiency, i.e., the ability to identify some samples belonging to it. In practice, this means that we need to specify different significance levels for the classes and for the global data set. To do so, we adapt the class-conditional Mondrian conformal prediction described in Section 2.

Let $\epsilon_g \in (0, 1)$ be the global error rate for all the data set, $\epsilon_0 \in (0, 1)$ be the one specified for the label $y = 0$, meaning that the person is not in debt, and $\epsilon_1 \in (0, 1)$ be the one related to the class $y = 1$, i.e. the person is in debt. ϵ_0 is therefore the variable over which we wish to have a strong control (in order not to send unnecessary social agents to the tenants). We have

$$\epsilon_g = \epsilon_0 \mathbb{P}(y = 0) + \epsilon_1 \mathbb{P}(y = 1). \tag{6}$$

With ϵ_g and ϵ_0 chosen by the user, and provided we have reasonable estimations of $\mathbb{P}(y = 0)$ and $\mathbb{P}(y = 1)$, we can calculate ϵ_1 by :

$$\epsilon_1 = \frac{\epsilon_g - \epsilon_0 \mathbb{P}(y = 0)}{\mathbb{P}(y = 1)}. \tag{7}$$

When defining ϵ_g and ϵ_0, and since $\epsilon_1 \in (0, 1)$, we need to respect the condition :

$$\epsilon_0 \mathbb{P}(y = 0) < \epsilon_g < \epsilon_0 + (1 - \epsilon_0)\mathbb{P}(y = 1), \tag{8}$$

otherwise we would obtain an unfeasible ϵ_1.

This enables us to have individual significance levels for each class that will guarantee an overall confidence level for the data set. Apart from this step, the other steps of class-conditional MCP remain the same. Thus, in our approach, we have the following procedure:

1. Split the original data set into a proper training set, a calibration set and a test set.
2. Use the proper training set to train the underlying algorithm, get the output predictions for the calibration and test sets and calculate their non-conformity scores based on the standard non-conformity measure in Equation (1).
3. Estimate the class prior probabilities $\mathbb{P}(y = 0)$ and $\mathbb{P}(y = 1)$ from the training set.
4. Fix ϵ_g and ϵ_0 and compute ϵ_1 based on the Equation (7), with respect to the specified condition in (8).
5. For each example in the test set, and for each class, compute the p-values using Equation (5) for class-conditional Mondrian conformal prediction.
6. For each example in the test set, get its set prediction using Equation (4).

Some comments are now in order. From Equation (7), it is clear that whenever ϵ_0 is fixed, the value of ϵ_1 increases as ϵ_g grows. Now, if our goal were to maximise our accuracy, we would set $\epsilon_g = \epsilon_0 = \epsilon_1$, as usual. However, in our setting, what is important is to identify a suitable number of persons that will be in debt rather than being too cautious. Hence, our goal for class 1 is to have a reasonable, if not maximal specificity, that is to ensure that a high amount of prediction sets (4) will contain only one value for class 1. As the size of Γ^ϵ decreases when ϵ increases, it is then desirable to have a high ϵ_1, hence to pick $\epsilon_g > \epsilon_0$.

4. Evaluation and discussion

In this section, we describe the experimental setting and the results of our study.

4.1. Data set

The data used in our paper is provided by Sopra Steria, a well-known French IT company that offers software for real estate rental management for its clients, mainly social property owners. The origin of our data set comes from a data warehouse of one of its clients that contains monthly historical records of tenants' activity from January 2018 to December 2019. This data was anonymized with respect to the General Data Protection Regulation (GDPR) EU law in order to protect the tenants' private data.

The data extraction procedure focused on information related to tenants, their personal situation (age, marital status, ...), their financial situation (job, salary, ...), their rental property (number of rooms, geographical location, ...), and payment transactions related to rent (rent, invoices, amounts collected, ...). Based on these monthly payment transactions, we added a new feature by computing the cumulative debt amount, which is a sum of the difference between the amount cashed and the invoice amount for each month. Hence, a person is considered in debt if their cumulative debt amount is greater than or equal to twice the gross monthly amount of their rent. Based on this value, we were able to add a boolean feature "in debt" which is equal to 0 if the person is not in debt, and 1 if they are. This feature will later be our class. Then, we created our examples by taking a period of three months prior to a debt occurrence (or randomly in case of no debt occurrence throughout the historical records), and added a feature "history start" to save the date of the first month of this time period. Note that for in debt people, and since for each example we take a history of 3 months on a period of almost 2 years, we can have more than one example for each person corresponding to different periods of time, in which this person can be in debt or not. This extraction strategy was chosen to have as much examples as possible and thus a bigger data set.

The data set used in this paper contains 28566 examples, 44 features and a binary class with 1 being in debt and 0 not being in debt. The data set is also strongly imbalanced as only 7.89% of the tenants are in debt, and has many missing values, due to the fact that some of the data is gathered through annual surveys. Due to the company's data and privacy policies, we unfortunately cannot share more information about the data set.

4.2. Experimental setting

In our study, we focused on two main experiments, the first one being a comparison between ICP and MCP with the same ϵ_g, and the second one being the adaptation of class-conditional MCP to control ϵ_g and ϵ_0 values.

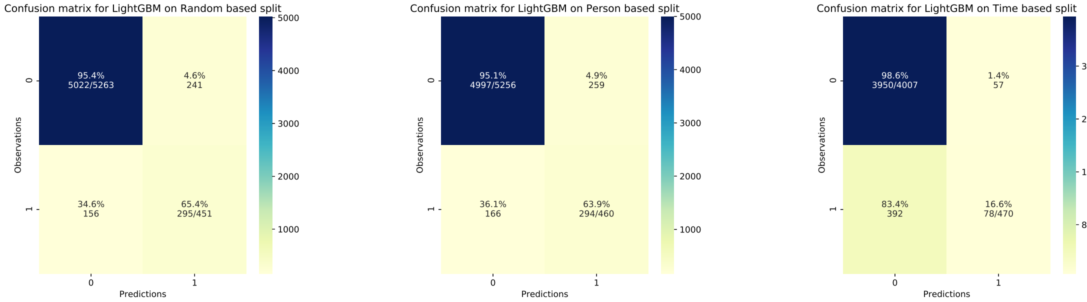
We chose the gradient boosting algorithm "LightGBM" as the underlying algorithm since it can handle missing values and both categorical and numerical features. We used the standard non-conformity measure $f(z) = 1 - \hat{P}_h[y \mid x]$. We also tried three types of splitting for the data set, with a test set equal to 20% of the data set, and with a calibration set equal to 20% of the training examples. These splitting approaches are as follows:

- **Random:** examples are split randomly on all training, calibration and test sets. From the application perspective, this is clearly unrealistic, as future events will be used to predict past ones (which is in practice impossible), and as the same tenant will possibly be in different sets, thus making unclear whether exchangeability holds. However, we would argue that this is true for most actual applications, where a task typically involves predicting future observations from past ones, and this does not stop most ML (including those on conformal prediction) papers to consider random splits on benchmarks to validate approaches. Since this is the standard splitting strategy that is used in all classic benchmarks, a random split therefore seems a good point to provide a proof-of-concept.

- **Person:** examples are split according to the tenant's ID, since we can have many examples for the same person. This means that the examples in the test set are those of people the algorithm did not see before in the training and calibration phases. This seems a more realistic scenario, as in practice tenants used in the training phase will often be past tenants, and tenants used in the operational phase will be new ones, different from the past ones.

- **Time:** the test predictions are done on examples from the future based on a training and calibration done on examples from the past. This is done by using randomised 2018 data for the training and calibration, and keeping 2019 data for the test based on the feature "history start", so that the test set was made of approximately 20% of the data set. This is undoubtedly the most realistic splitting strategy with respect to the application goal, however as we shall see this is also the one in which standard conformal prediction does not work very well, for potential reasons we will discuss.

For comparison purposes, Figure 1 shows the confusion matrix results for all splitting protocols with only the underlying algorithm ”LightGBM” with a thresh-hold of 0.3 (so as to settle a low percentage of tenants wrongly predicted as defaulting).



(a) Random (b) Person (c) Time

Figure 1: Confusion matrices for the underlying algorithm for all splitting protocols.

The first experiment was conducted based on the steps for the inductive conformal classifier as described in Section 2, for values of ϵ ranging from 0.1 to 0.9, where $\epsilon_g = \epsilon_0 = \epsilon_1$ in the case of MCP.

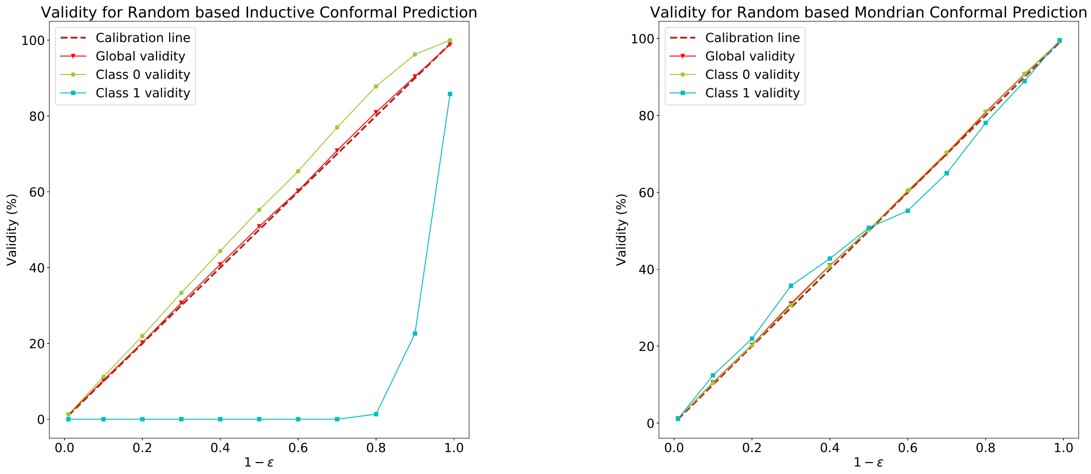
The second experiment was conducted by adapting class-conditional MCP to specify the error rate of the class 0 and the overall error rate, in order to limit the number of misclassified people that are predicted as in debt when in fact they are not. For this purpose, we followed our class-wise confidence approach as described in Section 3, with different ϵ_g and ϵ_0 values.

4.3. What did work

This section presents the promising results of our experiments, investigating in particular the difference between ICP and MCP, and the validity and efficiency of the proposed approach.

For our first experiment, and to verify the validity of ICP and MCP, we calculate the global accuracy and the accuracy of each class, and compare them with the calibration line. This line represents the case where the error rate is exactly equal to ϵ for a confidence level $1 - \epsilon$, which is what we seek to obtain in an exactly valid conformal predictor. Results are shown in Figures 2 and 3 for each Random and Person types of splitting.

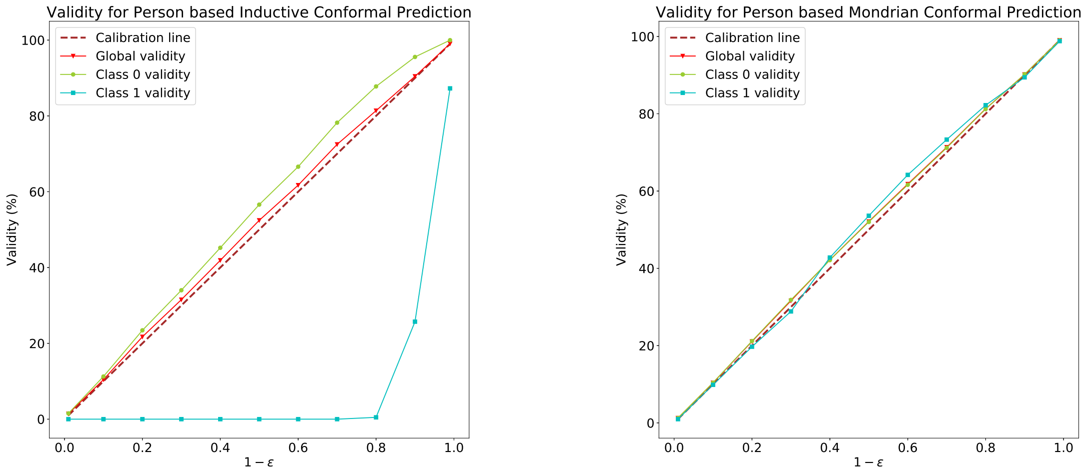
In the case of ICP (Figures 2(a) and 3(a)), results show that the global validity for all the data set is reached. However, it is not respected for the classes, more importantly for the class 1 corresponding to in debt tenants, since we have fewer examples for this class. This shows the problem of having an imbalanced data set in terms of categories or classes, and how the least represented area of the observations' space suffers the most when a simple conformal prediction method is used. Indeed, a very bad validity for the minority class can be compensated by a slightly conservative validity for the majority class. This problem is resolved in the case of MCP (Figures 2(b) and 3(b)), which gives better validity results that are almost exactly valid for the global data set and also for each individual class, including the minority class 1. In our case, a person-based split slightly outperforms a random-based split.



(a) ICP validity

(b) MCP validity

Figure 2: Validity results for Random-based split.



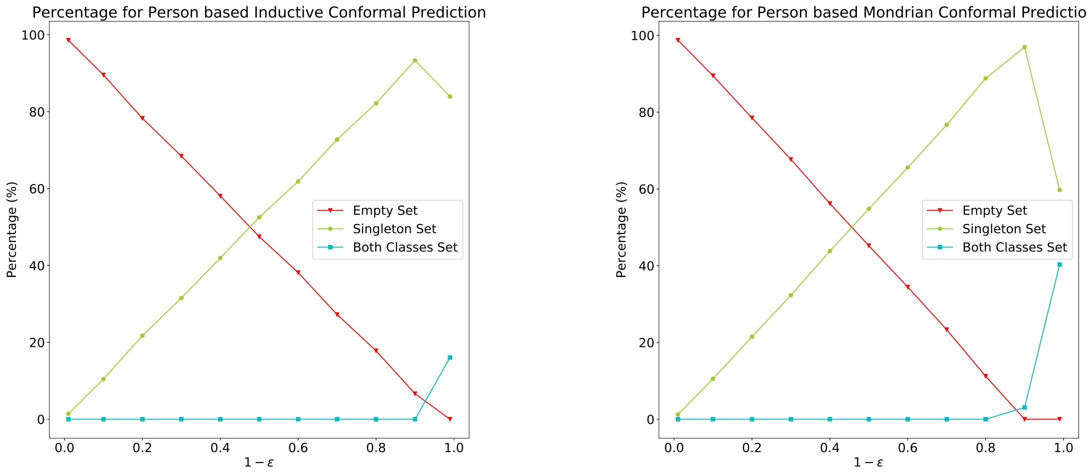
(a) ICP validity

(b) MCP validity

Figure 3: Validity results for Person-based split.

To evaluate the efficiency of ICP and MCP, we calculated the percentage of singletons, empty sets $\emptyset$ and $\{0, 1\}$ sets from all the predictions of the test examples. Results are shown in Figure 4 for person-based splits, and are similar for the random-based split.

For efficiency results, it is noticed that as ϵ decreases, the percentage of predicted empty sets $\emptyset$ lessens. It is even no longer predicted (at $\epsilon = 0.1$ for MCP). Conversely, the opposite



(a) ICP efficiency

(b) MCP efficiency

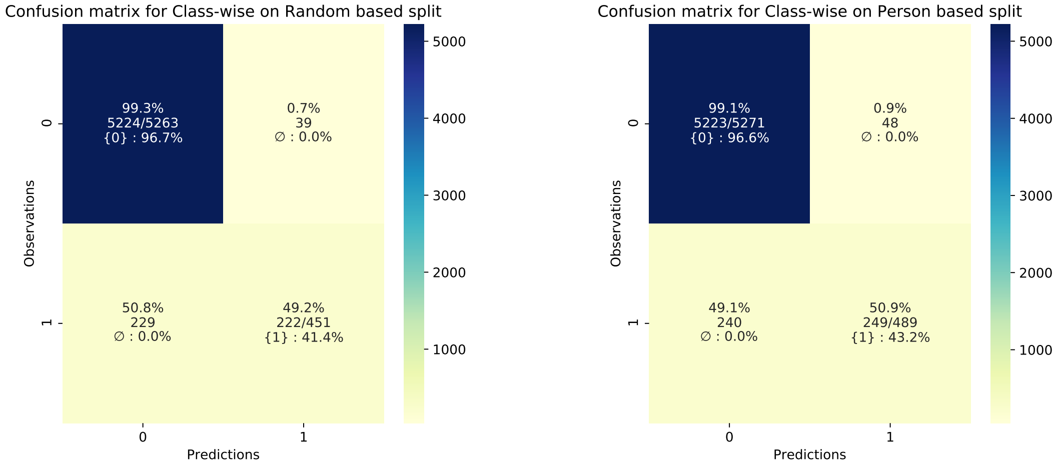
Figure 4: Percentage results for Person-based split.

is observed with the percentage of singleton sets which grows constantly as ϵ decreases until $\epsilon = 0.1$ for ICP and $\epsilon = 0.2$ for MCP. From that moment, we notice a mirror effect between the percentage of singletons and the percentage of $\{0, 1\}$ sets, which was null until now, and that grows whereas the percentage of singletons declines, with bigger values in the case of MCP. This can be explained by the fact that in MCP, the confidence level is guaranteed for each class as well as the global data set, meaning that the model predicts more $\{0, 1\}$ sets to have a more reliable prediction, even for the minority class. The same observations are made for both splitting methods.

For the second experiment, we used our class-wise confidence approach and specified different values for the significance levels ϵ_g and ϵ_0. Figures 5 and 6 show the confusion matrix for both random and person splits with $\epsilon_g = 0.05$ and $\epsilon_0 = 0.01$ and with $\epsilon_g = \epsilon_0 = 0.01$, corresponding to a classical MCP classifier.

In addition to having the percentage and amount of data falling in each cell, we have added an extra element, which is either the proportion of singleton predictions in the case of correct predictions (that is, the ratio of $\{0\}$ or $\{1\}$ among the predictions), or the proportion of empty set predictions in case of incorrect predictions.

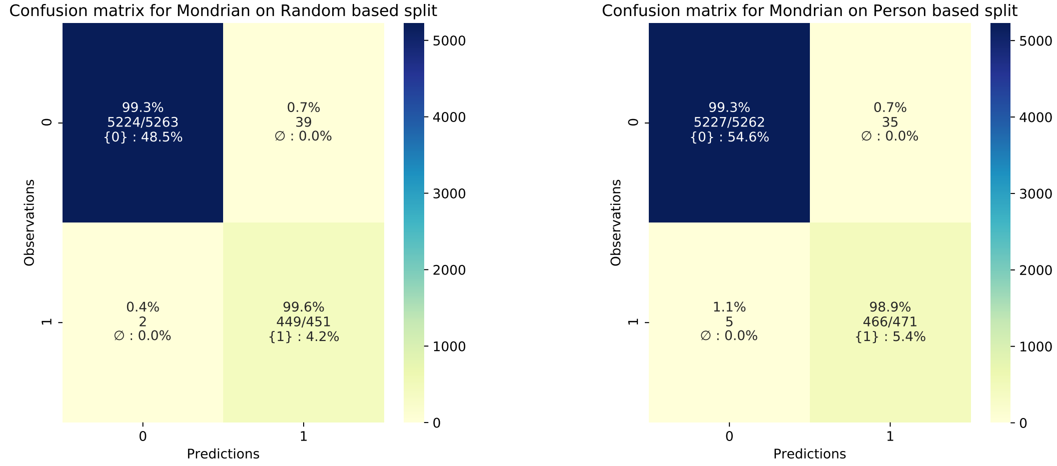
For our approach, Figure 5 shows that the error rate for class 0 is approximately equal to the chosen ϵ_0, and that the error rate for class 1 is also approximately 0.52, the result of Equation (7) when $\epsilon_0 = 0.01$ and $\epsilon_g = 0.05$. Also, the global validity is approximately equal to 95% as it was chosen by the social property owners, which shows that our method achieves a class-wise confidence while keeping a global confidence, both chosen by the user. Similarly, Figure 6 shows expected results for the more classical choice $\epsilon_g = \epsilon_0$. However, a striking difference between the two is the number of precisely recognised persons that will be in debt or not in debt, i.e. the percentage of singletons. For in debt persons, in the matrices of Figure 6, this amounts to 19 persons for the random-based split (4.2% of



(a) Random-based split

(b) Person-based split

Figure 5: Confusion matrix for class-wise confidence with $\epsilon_g = 0.05$ and $\epsilon_0 = 0.01$.



(a) Random-based split

(b) Person-based split

Figure 6: Confusion matrix for class-wise confidence with $\epsilon_g = \epsilon_0 = 0.01$.

449 persons) and 25 persons for the person-based split (5.4% of 466 persons), while in the matrices of Figure 5, this amounts to 91 persons for the random-based split (41.4% of 222 persons) and 108 persons for the person-based split (43.2% of 249 persons). So indeed our accuracy on the first class has dropped drastically in our scheme, even with a small margin between ϵ_g and ϵ_0, but the clear upside is that we were able to detect much more (about four times more) problematic tenants, allowing for more prevention. For people who are

not in debt, 2534 persons are precisely predicted in the case of classical MCP (Figure 6) for the random-based split (48.5% of 5224 persons) and 2854 persons for the person-based split (54.6% of 5227 persons), whereas with our class-wise approach (Figure 5), this amounts to 5052 persons for the random-based split (96.7% of 5224 persons) and 5056 persons for the person-based split (96.6% of 5233 persons), meaning that experts will have much less $\{0, 1\}$ cases to verify manually when using our method. This can be explained by the fact that, contrary to our approach, all ϵ values are equal in the classical MCP, thus $\epsilon_1 = 0.01$, which leads to more $\{0, 1\}$ sets predicted in order to ensure the 99% validity for the minority class. Consequently, this impacts the class 0, by decreasing the percentage of precisely predicted non in debt persons. Again, it is useful to recall that in this application, tenants that will be in debt and predicted as not in debt will anyway be detected and helped if possible. In Appendix A, Tables 1 and 2, we provide the full results obtained for various choices of ϵ_g and ϵ_0, in the case of random and person splits.

As a conclusion, the first feedback from Sopra Steria and its client (the social property owner) was enthusiastic and showed the advantages of using our class-wise approach, especially when it comes to the absolute number of well-verified tenants, a more important indicator for our data provider than the ratio. Indeed, using the person splitting strategy, it is possible to reserve an amount of tenants' data for training and calibration (for instance, tenants that are no longer customers of the social property owner), and only predict in the production phase on new never-seen-before tenants. Our empirical results also tend to indicate that the common assumption of conformal prediction (variable exchangeability, i.e., the fact that the drawn inferences are invariable under observation permutation) holds at least approximately in those cases.

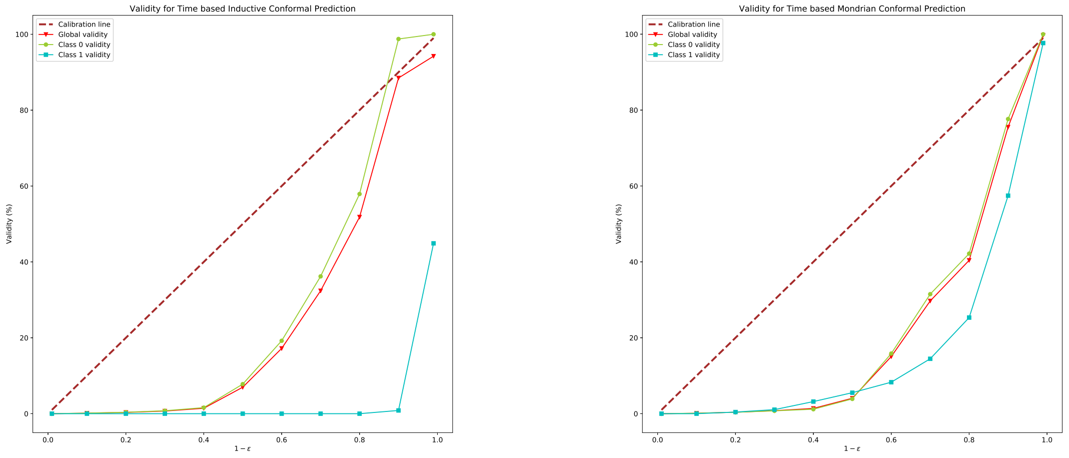
They should also be compared to the confusion matrices of Figure 1, where the number of tenants falsely predicted as defaulting is as high as the number of tenants rightly predicted as defaulting, both for the random and person splits. This also shows that the conformal approaches can bring a significant edge when compared to standard methods.

4.4. What did not work

This section presents the results of our experiments for the time splitting strategy. For our first experiment, we verify the validity of ICP and MCP for the time splitting strategy by comparing the global accuracy and the accuracy of each class with the calibration line. Results are shown in Figure 7.

In both cases, we notice that both methods are performing actually very badly, with the global and class accuracies below the calibration line. We also observe slightly better results for the MCP when it comes to the minority class 1, with its accuracy being closer to the global and majority class ones as compared to the ICP. We will discuss the possible reasons behind such a result in the next section.

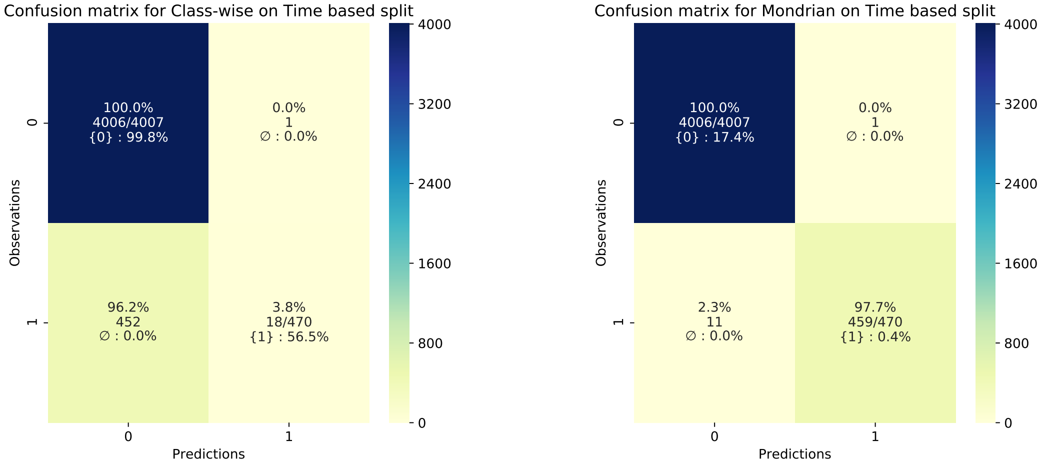
For the second experiment, Figure 8 shows the confusion matrix of our class-wise approach for time split with $\epsilon_g = 0.05$ and $\epsilon_0 = 0.01$ and with $\epsilon_g = \epsilon_0 = 0.01$, corresponding to a Mondrian approach. These results go together with the results of our first experiment, showing that the chosen ϵ values are not respected for the class-wise approach with an actual global accuracy of 89.88% instead of 95% and an actual accuracy of 3.83% instead of the computed 45% for class 0 (c.f. Appendix A Table 3 for full results). When closely



(a) ICP validity

(b) MCP validity

Figure 7: Validity results for Time-based split.



(a) $\epsilon_g = 0.05$ and $\epsilon_0 = 0.01$.

(b) $\epsilon_g = \epsilon_0 = 0.01$.

Figure 8: Confusion matrix for class-wise confidence with a Time-based split.

examining the results, we also observe that for MCP (Figure 8(b)), the percentage of singletons is very small for all of the confusion matrix cells, meaning that the experts should manually verify nearly all examples. The only case where the amount of singleton is high (Figure 8(a)) is when most predictions belong to the class 0, in which case they are totally uninformative (hence useless).

As a conclusion, the time-based strategy, the splitting protocol that uses the temporal aspect of the data set and is the most consistent with our final applied intent, did not work. We discuss the possible fixes to that in the next section.

4.5. What's next

Overall, some of our experiments have shown that conformal approaches can be used to finely control class-wise defined errors, the need of which may easily arise in applications where some type of errors should be controlled finely, while other ones are deemed less important.

However, not everything worked, and even in those cases where it worked, there are probably still some room for improvement:

- **Time-based split:** in this case, the lack of an exact validity could be explained by the fact that the time splitting strategy violates the exchangeability assumption, possibly by a change of conditions not detected and not present in the data that changes the underlying distribution. Discussions with our end-user (Sopra Steria) did not make us able to identify a possible source for such a change, as the data set is a real, but outdated one. One possibility to solve this issue would be to explore, use and/or adapt conformal approaches able to deal with (temporal) distribution shift in the test set (Gibbs and Candès (2021)).

- **Better control of the error to satisfy the end-user:** what Mondrian conformal prediction can provide is a control of the class-conditional error, that is, knowing that $y = c$, we have validity on the probability $P(y \in \hat{Y} | y = c)$ where $\hat{Y}$ is our conformal prediction. However, what the user is ultimately interested in is the control of $P(y = c | \hat{Y} = \{c\})$, of which our approach only provides a proxy (deemed satisfying by the user). Our next step would therefore be to search to control $P(y = c | \hat{Y} = \{c\})$, possibly by combining different conformal predictors (e.g., Class-wise with global ones), or by using Venn-Abers predictors (Vovk and Petej (2014)) that are able to provide calibrated probabilities as outputs, with the caveats that they provide multiple values for those, of which at least one is ensured to be calibrated.

5. Conclusion

In this paper, we used the conformal prediction, and in particular the class-conditional Mondrian variant, to have more control on the error rate of a certain class, while preserving an overall confidence. We applied our method to the real estate domain in order to help social property owners limit the number of people that are falsely predicted as in debt when, in fact, they are not, as these misclassifications are costly. The results show the interest of this method on our data set when two types of splits are considered (random and person based), with less-satisfying results when using the time-based split.

From a methodological perspective, our contribution is rather modest but opens up questions about a problem we think is important: the one of identifying how various error rates should be elicited/chosen when considering conformal frameworks with such multiple error rates. Indeed, while choosing equal error rates is common in systematic studies, it can hardly be expected in real-life applications that all targets need the same accuracy. This applies of course in frameworks where Mondrian conformal prediction is at play, but also in multi-target settings such as multi-variate regression (Messoudi et al. (2020); Neeven and Smirnov (2018)) or multi-label classification (Hüllermeier et al. (2020); Lambrou and Papadopoulos (2016)). Extending our current approach to more than two classes in the Mondrian case would require considering the constraints connecting the different confidence degrees, adapting Equation (8) to the multi-class case. The multi-target setting would be less problematic, as confidence degrees over the different targets are not constrained by one another. It would, however, require to be able to construct the link between the global confidence degree and each individual confidence degree, using for instance copula-based frameworks (Messoudi et al. (2021)).

Concerning our application, the perspectives include treating missing values to improve classification results, as for the moment we use a model that handles them naturally. Also, we would like to explore other non-conformity measures other than the standard one used in this paper. Moreover, it would be interesting to work on the time-based split, by applying conformal prediction methods that treat temporal data such as time series (Chernozhukov et al. (2018)), or by considering the combination of conformal approaches with robust approaches to the distribution shift in order to obtain validity even when the data generating distribution varies over time.

References

Emmanuel J Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*, 2021.

Victor Chernozhukov, Kaspar Wüthrich, and Zhu Yinchu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On Learning Theory*, pages 732–749. PMLR, 2018.

Dmitry Devetyarov, Ilia Nouretdinov, Brian Burford, Stephane Camuzeaux, Aleksandra Gentry-Maharaj, Ali Tiss, Celia Smith, Zhiyuan Luo, Alexey Chervonenkis, Rachel Hallett, et al. Conformal predictors in early diagnostics of ovarian and breast cancers. *Progress in Artificial Intelligence*, 1(3):245–257, 2012.

Alex Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, page 148–155, 1998.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. *arXiv preprint arXiv:2106.00170*, 2021.

Eyke Hüllermeier, Johannes Fürnkranz, and Eneldo Loza Mencia. Conformal rule-based multi-label classification. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 290–296. Springer, 2020.

Antonis Lambrou and Harris Papadopoulos. Binary relevance multi-label conformal predictor. In *Symposium on Conformal and Probabilistic Prediction wIth Applications*, pages 90–104. Springer, 2016.

Michael Manville, Paavo Monkkonen, Michael Lens, and Richard Green. Covid-19 and renter distress: Evidence from los angeles. 2020.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Conformal multi-target regression using neural networks. In *Conformal and Probabilistic Prediction and Applications*, pages 65–83. PMLR, 2020.

Soundouss Messoudi, Sébastien Destercke, and Sylvain Rousseau. Copula-based conformal prediction for multi-target regression. *Pattern Recognition*, Accepted for publication, 2021.

Jelmer Neeven and Evgueni Smirnov. Conformal stacked weather forecasting. In *Conformal and Probabilistic Prediction and Applications*, pages 220–233. PMLR, 2018.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. IntechOpen, 2008.

Paolo Toccaceli and Alexander Gammerman. Combination of inductive mondrian conformal predictors. *Machine Learning*, 108(3):489–510, 2019.

Vladimir Vovk and Ivan Petej. Venn-abers predictors. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 829–838, 2014.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Fan Yang, Hua-zhen Wang, Hong Mi, Wei-wen Cai, et al. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC bioinformatics*, 10 (1):1–14, 2009.

Shiwei Zhang, Xiuzhen Zhang, Jey Han Lau, Jeffrey Chan, and Cecile Paris. Less is more: Rejecting unreliable reviews for product question answering. *arXiv preprint arXiv:2007.04526*, 2020.

Appendix A. Additional results

Tables 1, 2 and 3 present a complete overview of our results for different choices of ϵ_0 and ϵ_g for all splitting strategies.

Epsilons			Accuracy (%)			$\{0,1\}$ sets (%)		$\emptyset$ sets (%)	
ϵ_g (chosen)	ϵ_0 (chosen)	ϵ_1 (computed)	Global	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
0.01	0.01	0.01	99.28	99.26	99.56	51.5	95.81	0.0	0.0
0.02	0.01	0.14	98.04	99.26	83.81	14.15	86.11	0.0	0.0
0.03	0.01	0.26	97.22	99.26	73.39	8.91	79.47	0.0	0.0
0.04	0.01	0.39	95.89	99.26	56.54	4.95	67.95	0.0	0.0
0.05	0.01	0.52	95.31	99.26	49.22	3.33	58.61	0.0	0.0
0.06	0.01	0.64	94.56	99.26	39.69	1.91	44.69	0.0	0.0
0.07	0.01	0.77	93.45	99.26	25.72	0.0	0.0	5.13	2.69
0.08	0.01	0.9	92.4	99.26	12.42	0.0	0.0	69.23	17.47
0.05	0.05	0.05	95.17	95.15	95.34	18.7	76.33	0.0	0.0
0.06	0.05	0.18	93.93	95.15	79.6	5.39	47.81	0.0	0.0
0.07	0.05	0.3	92.72	95.15	64.3	0.0	0.0	10.2	4.97
0.08	0.05	0.43	91.9	95.15	53.88	0.0	0.0	39.22	26.44
0.09	0.05	0.56	91.23	95.15	45.45	0.0	0.0	60.0	37.8
0.1	0.05	0.68	90.58	95.15	37.25	0.0	0.0	73.73	45.94
0.11	0.05	0.81	89.24	95.15	20.18	0.0	0.0	89.02	57.5
0.12	0.05	0.94	88.26	95.15	7.76	0.0	0.0	96.86	63.22
0.1	0.1	0.1	90.71	90.86	88.91	5.18	40.98	0.0	0.0
0.11	0.1	0.23	89.66	90.86	75.61	0.0	0.0	22.87	17.27
0.12	0.1	0.35	88.41	90.86	59.87	0.0	0.0	58.84	49.72
0.13	0.1	0.48	87.78	90.86	51.88	0.0	0.0	72.97	58.06
0.14	0.1	0.61	87.03	90.86	42.35	0.0	0.0	81.08	65.0
0.15	0.1	0.73	86.09	90.86	30.38	0.0	0.0	90.23	71.02
0.16	0.1	0.86	85.11	90.86	17.96	0.0	0.0	96.05	75.41
0.14	0.15	0.02	87.03	86.03	98.67	35.3	79.97	0.0	0.0
0.15	0.15	0.15	85.74	86.03	82.26	0.0	0.0	28.84	41.25
0.16	0.15	0.28	84.65	86.03	68.51	0.0	0.0	62.18	66.9
0.17	0.15	0.4	83.6	86.03	55.21	0.0	0.0	77.14	76.73
0.18	0.15	0.53	83.13	86.03	49.22	0.0	0.0	83.81	79.48
0.19	0.15	0.66	82.32	86.03	39.02	0.0	0.0	89.39	82.91
0.2	0.15	0.78	81.06	86.03	23.06	0.0	0.0	95.65	86.46
0.21	0.15	0.91	80.08	86.03	10.64	0.0	0.0	98.78	88.34
0.19	0.2	0.07	81.85	81.0	91.8	0.0	0.0	17.8	37.84
0.2	0.2	0.2	80.77	81.0	78.05	0.0	0.0	58.3	76.77
0.21	0.2	0.33	79.47	81.0	61.64	0.0	0.0	79.1	86.71
0.22	0.2	0.45	78.74	81.0	52.33	0.0	0.0	86.2	89.3
0.23	0.2	0.58	78.11	81.0	44.35	0.0	0.0	90.3	90.84
0.24	0.2	0.71	77.28	81.0	33.92	0.0	0.0	94.2	92.28
0.25	0.2	0.83	76.13	81.0	19.29	0.0	0.0	97.5	93.68
0.26	0.2	0.96	74.99	81.0	4.88	0.0	0.0	99.7	94.64
0.24	0.25	0.12	76.86	76.1	85.81	0.0	0.0	52.7	79.69
0.25	0.25	0.25	75.95	76.1	74.28	0.0	0.0	72.66	88.79
0.26	0.25	0.38	74.57	76.1	56.76	0.0	0.0	85.93	93.33
0.27	0.25	0.5	74.08	76.1	50.55	0.0	0.0	90.14	94.17
0.28	0.25	0.63	73.29	76.1	40.58	0.0	0.0	92.93	95.15
0.29	0.25	0.76	72.24	76.1	27.27	0.0	0.0	96.98	96.04
0.3	0.25	0.88	71.21	76.1	14.19	0.0	0.0	98.97	96.64

Table 1: Summary results for class-wise confidence with different values of ϵ for Random split.

	Epsilons		Accuracy (%)			$\{0,1\}$ sets (%)		$\emptyset$ sets (%)	
ϵ_g (chosen)	ϵ_0 (chosen)	ϵ_1 (computed)	Global	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
0.01	0.01	0.01	99.3	99.33	98.94	45.44	94.63	0.0	0.0
0.02	0.01	0.13	98.37	99.06	90.5	17.38	88.92	0.0	0.0
0.03	0.01	0.26	97.29	99.06	77.11	9.98	82.01	0.0	0.0
0.04	0.01	0.38	96.09	99.06	62.2	5.93	72.79	0.0	0.0
0.05	0.01	0.54	95.0	99.09	50.92	3.37	56.78	0.0	0.0
0.06	0.01	0.63	94.21	99.06	38.88	1.74	43.4	0.0	0.0
0.07	0.01	0.75	93.0	99.06	23.76	0.0	0.0	16.0	2.83
0.08	0.01	0.88	92.25	99.06	14.47	0.0	0.0	64.0	13.38
0.05	0.05	0.05	94.32	94.43	93.04	16.67	72.1	0.0	0.0
0.06	0.05	0.18	93.32	94.43	80.65	4.13	38.76	0.0	0.0
0.07	0.05	0.31	92.29	94.43	67.83	0.0	0.0	12.63	10.14
0.08	0.05	0.44	90.91	94.43	50.65	0.0	0.0	46.08	41.41
0.09	0.05	0.57	89.79	94.43	36.74	0.0	0.0	68.6	54.3
0.1	0.05	0.7	89.07	94.43	27.83	0.0	0.0	82.25	59.94
0.11	0.05	0.83	88.37	94.43	19.13	0.0	0.0	90.78	64.25
0.12	0.05	0.95	87.24	94.43	5.0	0.0	0.0	97.95	69.57
0.1	0.1	0.1	89.93	90.0	89.02	2.7	25.81	0.0	0.0
0.11	0.1	0.23	88.64	90.0	71.96	0.0	0.0	44.95	50.0
0.12	0.1	0.36	87.78	90.0	60.51	0.0	0.0	62.1	64.5
0.13	0.1	0.49	86.64	90.0	45.33	0.0	0.0	79.05	74.36
0.14	0.1	0.62	85.9	90.0	35.51	0.0	0.0	87.62	78.26
0.15	0.1	0.75	85.02	90.0	23.83	0.0	0.0	95.24	81.6
0.16	0.1	0.88	84.0	90.0	10.28	0.0	0.0	97.52	84.38
0.14	0.15	0.02	86.18	85.08	98.31	22.78	70.41	0.0	0.0
0.15	0.15	0.15	85.03	85.08	84.39	0.0	0.0	28.79	43.24
0.16	0.15	0.28	84.29	85.08	75.53	0.0	0.0	60.38	63.79
0.17	0.15	0.41	82.9	85.08	58.65	0.0	0.0	77.45	78.57
0.18	0.15	0.54	81.85	85.08	45.99	0.0	0.0	85.61	83.59
0.19	0.15	0.68	80.77	85.08	32.91	0.0	0.0	93.12	86.79
0.2	0.15	0.81	79.68	85.08	19.62	0.0	0.0	97.71	88.98
0.21	0.15	0.94	78.56	85.08	6.12	0.0	0.0	99.49	90.56
0.19	0.2	0.08	80.39	79.47	91.15	0.0	0.0	20.35	32.5
0.2	0.2	0.2	79.69	79.47	82.3	0.0	0.0	58.66	66.25
0.21	0.2	0.32	78.89	79.47	72.12	0.0	0.0	72.65	78.57
0.22	0.2	0.44	78.19	79.47	63.27	0.0	0.0	81.22	83.73
0.23	0.2	0.55	77.22	79.47	50.88	0.0	0.0	88.58	87.84
0.24	0.2	0.67	76.47	79.47	41.37	0.0	0.0	92.91	89.81
0.25	0.2	0.79	75.49	79.47	28.98	0.0	0.0	95.86	91.59
0.26	0.2	0.91	74.17	79.47	12.17	0.0	0.0	98.8	93.2
0.24	0.25	0.12	76.37	75.47	87.0	0.0	0.0	54.07	77.59
0.25	0.25	0.25	75.34	75.47	73.77	0.0	0.0	76.73	88.89
0.26	0.25	0.38	74.53	75.47	63.45	0.0	0.0	84.33	92.02
0.27	0.25	0.51	73.43	75.47	49.33	0.0	0.0	90.15	94.25
0.28	0.25	0.64	72.34	75.47	35.43	0.0	0.0	94.57	95.49
0.29	0.25	0.77	71.18	75.47	20.63	0.0	0.0	97.52	96.33
0.3	0.25	0.89	70.29	75.47	9.19	0.0	0.0	98.91	96.79

Table 2: Summary results for class-wise confidence with different values of ϵ for Person split.

Epsilons			Accuracy (%)			$\{0, 1\}$ sets (%)		$\emptyset$ sets (%)	
ϵ_g (chosen)	ϵ_0 (chosen)	ϵ_1 (computed)	Global	Class 0	Class 1	Class 0	Class 1	Class 0	Class 1
0.01	0.01	0.01	99.73	99.98	97.66	82.61	99.65	0.0	0.0
0.02	0.01	0.14	93.63	99.98	39.57	18.59	98.35	0.0	0.0
0.03	0.01	0.28	91.07	99.98	15.11	2.58	88.98	0.0	0.0
0.04	0.01	0.41	90.24	99.98	7.23	1.14	77.97	0.0	0.0
0.05	0.01	0.55	89.88	99.98	3.83	0.25	43.48	0.0	0.0
0.06	0.01	0.68	89.66	99.98	1.7	0.0	0.0	100.0	1.08
0.07	0.01	0.82	89.5	99.98	0.21	0.0	0.0	100.0	2.56
0.05	0.05	0.05	96.65	98.7	79.15	46.45	96.44	0.0	0.0
0.06	0.05	0.18	91.29	98.7	28.09	6.2	77.33	0.0	0.0
0.07	0.05	0.32	89.75	98.7	13.4	0.0	0.0	17.31	2.46
0.08	0.05	0.45	89.01	98.7	6.38	0.0	0.0	63.46	9.77
0.09	0.05	0.59	88.7	98.7	3.4	0.0	0.0	92.31	12.56
0.1	0.05	0.72	88.43	98.7	0.85	0.0	0.0	100.0	14.81
0.11	0.05	0.86	88.36	98.7	0.21	0.0	0.0	100.0	15.35
0.1	0.1	0.1	75.52	77.64	57.45	6.76	48.08	0.0	0.0
0.11	0.1	0.23	71.57	77.64	19.79	0.0	0.0	90.62	36.34
0.12	0.1	0.37	70.45	77.64	9.15	0.0	0.0	96.32	43.79
0.13	0.1	0.5	70.05	77.64	5.32	0.0	0.0	98.44	46.07
0.14	0.1	0.64	69.78	77.64	2.77	0.0	0.0	99.89	47.48
0.15	0.1	0.77	69.53	77.64	0.43	0.0	0.0	100.0	48.72
0.14	0.15	0.02	57.49	52.93	96.38	28.46	60.81	0.0	0.0
0.15	0.15	0.15	51.46	52.93	38.94	0.0	0.0	68.08	75.26
0.16	0.15	0.28	48.89	52.93	14.47	0.0	0.0	97.51	82.34
0.17	0.15	0.42	48.13	52.93	7.23	0.0	0.0	98.62	83.72
0.18	0.15	0.55	47.76	52.93	3.62	0.0	0.0	99.68	84.33
0.19	0.15	0.69	47.51	52.93	1.28	0.0	0.0	100.0	84.7
0.2	0.15	0.82	47.4	52.93	0.21	0.0	0.0	100.0	84.86
0.19	0.2	0.07	44.9	42.18	68.09	0.0	0.0	45.23	86.0
0.2	0.2	0.2	40.41	42.18	25.32	0.0	0.0	93.01	94.02
0.21	0.2	0.33	39.04	42.18	12.34	0.0	0.0	98.19	94.9
0.22	0.2	0.47	38.4	42.18	6.17	0.0	0.0	99.31	95.24
0.23	0.2	0.6	38.08	42.18	3.19	0.0	0.0	99.87	95.38
0.24	0.2	0.74	37.84	42.18	0.85	0.0	0.0	100.0	95.49
0.25	0.2	0.87	37.77	42.18	0.21	0.0	0.0	100.0	95.52
0.24	0.25	0.12	37.06	35.59	49.57	0.0	0.0	65.01	93.25
0.25	0.25	0.25	33.71	35.59	17.66	0.0	0.0	97.21	95.87
0.26	0.25	0.38	32.79	35.59	8.94	0.0	0.0	98.92	96.26
0.27	0.25	0.52	32.32	35.59	4.47	0.0	0.0	99.65	96.44
0.28	0.25	0.65	32.1	35.59	2.34	0.0	0.0	100.0	96.51
0.29	0.25	0.79	31.9	35.59	0.43	0.0	0.0	100.0	96.58

Table 3: Summary results for class-wise confidence with different values of ϵ for Time split.