ML4DS Final Report - Group G

February 11, 2024

André Hammernik (HSB) Dominik Czajka (Vistula) Gustavo Guedes (IPB) Malte-Sweer Schubert (HSB)

1 Problem Description

As part of blended intensive programmes (BIP), the task was to use a suitable machine learning algorithm based on a dataset and to classify a competition dataset correctly. The project involves different groups consisting of 5 students from different countries. The aim is, among other things, a small competition among each other regarding the accuracy of the chosen machine learning algorithm. Each team was allowed to submit 3 classifications. The competition dataset does not contain the label consumer type and serves as a comparison for the predictive power of the machine learning methods between the groups. The following paragraph describes the basic characteristics of the dataset. In particular, only a subset of the characteristics is discussed in order to make the report concise.

As part of the preparation for the personal group work, the group members independently dealt with data understanding, presented interesting representations within the group and visualized possible correlations. Data Understanding required the most time and therefore served as good preparation for the weeks before the stay in Braganca. First and foremost, an understanding of the data for the training dataset was gained, but characteristics of the competition dataset were also identified. The most interesting aspects for us are described below.

The dataset contains entries over a period from January 2013 to December 2020, whereby the year 2015 is completely missing. There are 7 different consumer types and 49 different installation zones. Based on the consumer number, it can be determined that the dataset contains 27,632 different consumers.

Table 1 shows how many entries there are for each consumer type. It is clearly recognisable that the data is very unbalanced. The domestic type is by far the most represented. Rural domestic is the second largest group with around a quarter of households compared to domestic. This is followed by industrial, which has significantly more data than rural commercial. Low income families and rural expansion bring up the rear with fewer than 1000 entries each.

consumer type	count
domestic	236,167
rural domestic	63,086
industrial	21,057
rural commercial	5,541
construction	2,235
low income families	999
rural expansion	890

Table 1: count of entries per consumer type

By looking at the distribution of the data over the years (see figure 1), it can be seen that they are distributed relatively evenly, although there is a slight increase over the years. Again, it can be seen that the year 2015 is completely missing. The average consumption in Figure 2 shows no positive or negative trend and fluctuates only slightly 0.3 units over the years. The average consumption fluctuates around 6.7 units. Starting at approximately 6.81 in 2013, it decreases to 6.63 in 2016, peaks at 6.88 in 2017, hits a minimum around 2019 with 6.56, and then rises again to 6.8 in 2020.

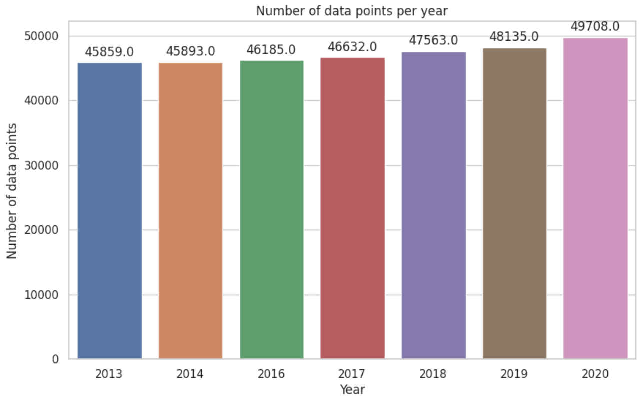


Figure 1: Data Points Per Year

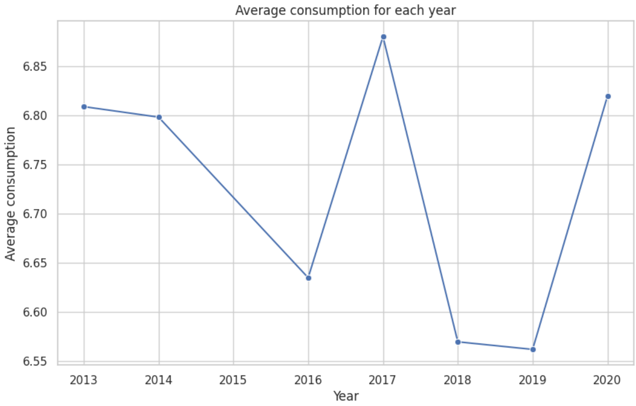


Figure 2: Average Consumption

In order to analyse whether there are seasonal fluctuations in consumption between the months, the line diagram from Figure 3 is generated. This shows a line with the average consumption per month for each year. It can be seen that there is indeed a seasonal difference as assumed, as more water is consumed in summer than in winter.

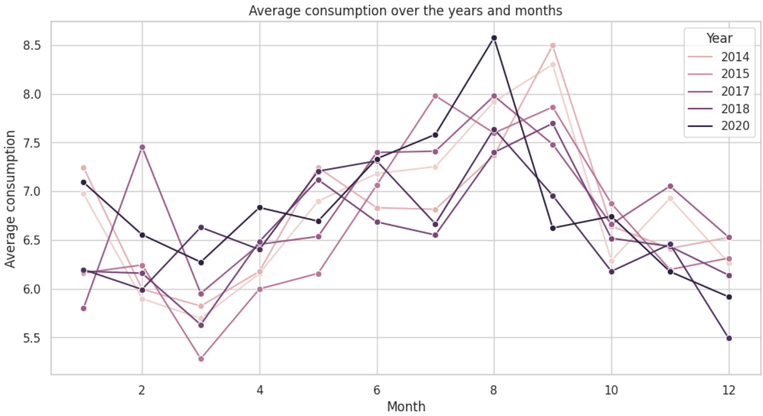


Figure 3: average consumption per year

Figure 4 shows the average consumption per consumer type, which fluctuates a little and ranges between approximately 3.5 and 11. The attempt to also consider the quartiles of consumption led to the box plot in Figure 5. Reference was again made to the individual years. However, no boxes can be recognised as the consumption scale goes up to 5,000 because some values are far above the average. There are some outliers in the dataset that still need to be removed during pre-processing.

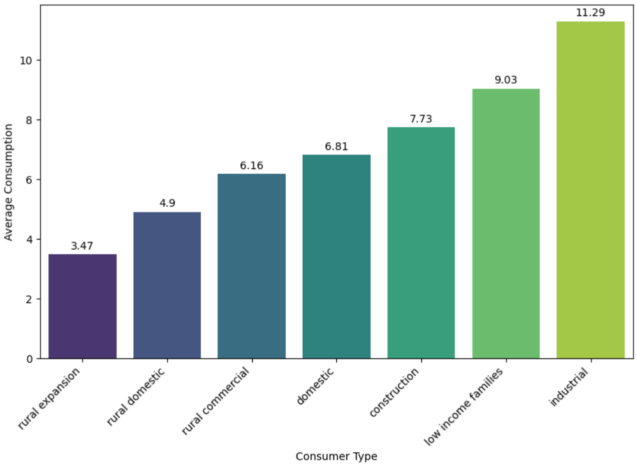


Figure 4: consumption per consumer type

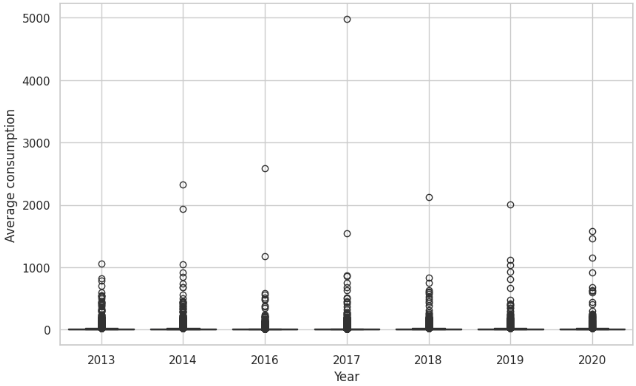


Figure 5: Boxplot

Using the label encoder, the labels of Consumer_type and Installation_zone were converted into numerical values in order to set up a correlation matrix on the uncleaned data (see Figure 6) . Using this correlation matrix, it was found that there is a strong relationship between the Installation_zone and the Consumer_type.

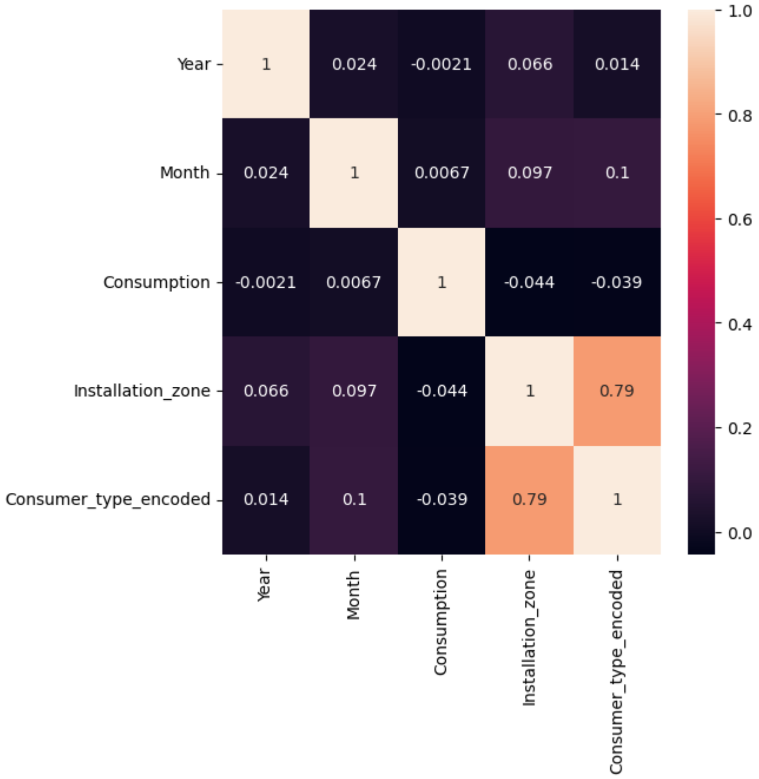


Figure 6: Correlation Matrix

Based on this, the frequency with which a consumer_type occurs in a specific installation_zone was analysed. Table 2 shows an excerpt from the resulting table. Considering the geographical distribution by incorporating 'installation_zone', it is observed that most domestic entries (73,358) are in Installation Zone 1. Installation Zone 1 also has a significant representation of industrial entries with 3,511. Another prominent area is Installation Zone 2, contributing 41,103 entries for Domestic and 6,160 for Industrial. Installation Zones 3 and 4 together have an additional 121,676 entries for Domestic. Thus, these four zones cover almost all Domestic entries, leaving only 30 entries.

	construction	domestic	industrial	low income	rural	rural	rural
Installation_zone				families	commercial	domestic	expansion
1	230	73,358	3,511	322	51	0	0
2	384	41,103	6,160	113	1,943	0	11
3	733	58,190	5,197	368	1,800	23	1
4	386	63,486	6,020	171	0	18	0
5	33	0	16	0	0	2,518	20
6	20	0	0	0	0	1,093	53
...	...	...	...	...	...	...	...

Table 2: number of entries per installation zone and consumer type (incomplete)

Taking a closer look at consumer numbers, it is found that they consist of 4 letters and 14 digits. The dataset includes 27,632 different consumer numbers, with only 307 different ones in construction, but 18,278 different ones in domestic.

For each ConsumerNumber in construction, there is an average of 7.2 entries, while domestic has an average of 12.9 entries. Low-income families have the lowest average entries with only 6.3 per unique consumer number.

2 Methodology

The following paragraph will describe our data pre-processing and our approaches, which we focused on the most, our metrics used, and our approach to evaluation to select the most appropriate approach.

2.1 Data preprocessing

One of the basic steps preceding any analysis and interpretation of data is its proper preparation. We are talking primarily about locating the data relevant to the planned analytical activities, obtaining them in the most convenient form possible and transformation, which may involve several more or less complex steps. How is data preparation defined? In our opinion data preparation should be understood as any activity undertaken to improve the quality, usability or availability of data, including, but not limited to: integration, profiling, cleaning and data management.

Based on the previously identified aspects, we agreed on different ways of preparing the data. The simplest one was to remove duplicate entries. We made sure that all columns were identical before deleting the entry. This looked like a duplication of entries to us and offered no added value.

In addition, we classified the consumption of 0 as an anomaly, as a consumption of 0 did not seem to make sense to us. It could be the result of other data processing, how incorrect values were handled, the result of rounding or other reasons.

We also looked at the outliers using the boxplot and decided that we would consider values above the 1.5 IQR as outliers. After removing all entries above the calculated value for the consumer type, the boxplot (Figure 7) looked much more meaningful. Despite the removal of outliers, it was not possible to achieve an ideal separation in the boxplots of the individual customer types.

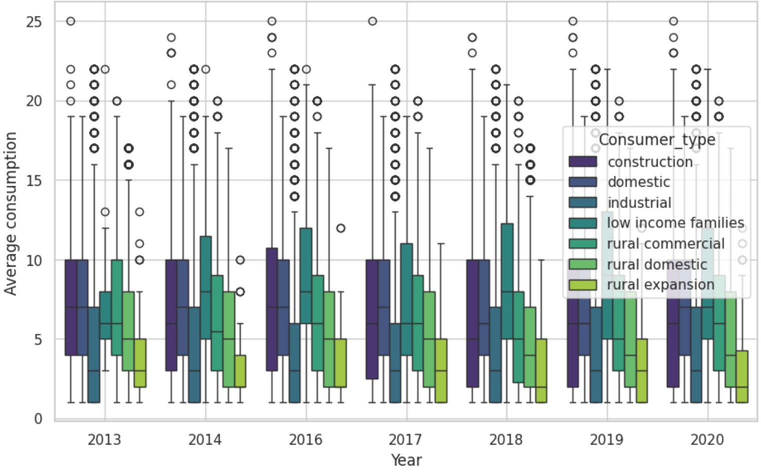
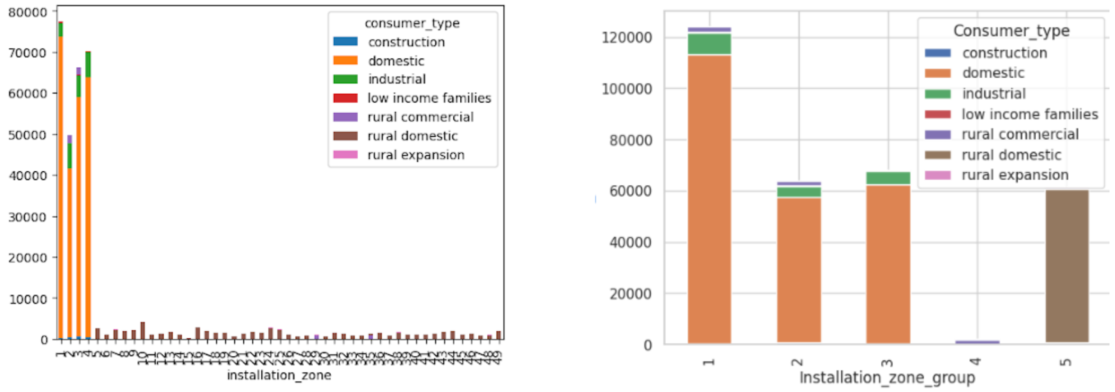


Figure 7: Boxplot after cleaning

Another point that we looked at more closely was the installation zone. With 49 different ones, there were too much verity to keep track. We discovered when we generated a simple decision tree with the unprocessed data that the individual installation zones were separated and when looking at the installation zones we could already recognize certain characteristics and similarities. Looking deeper, we looked at the consumer type assignment to the individual zones. We decided to put zones 1 and 2 together, left 3 and 4 as they were, put 29 and 35 together and put all the remaining ones in a new additional group. This reduces the distribution from 49 to 5 different zones.



(a) before grouping

(b) after grouping

Figure 8: Comparison of Distribution of Consumer types in installationzones before and after grouping

We discussed how to deal with the missing year 2015. As we didn't work directly with time series, we decided that we didn't need to do any feature engineering for it. We preferred to focus the short time on other aspects.
As we noticed in 1 the dataset is hugely imbalanced, our elephant in the room! The essence of the imbalance problem is that applying classical learning mechanisms on an unbalanced dataset can lead to the learned classifier favoring the dominant class at the expense of the dominated class - in this case 'domestic'. All classes except for 'domestic' and 'rural domestic' have been marginalized. It was leading to accuracy for these specific classes on level between 0-3 percent. What we decided to do was to test different ways of balancing our data.

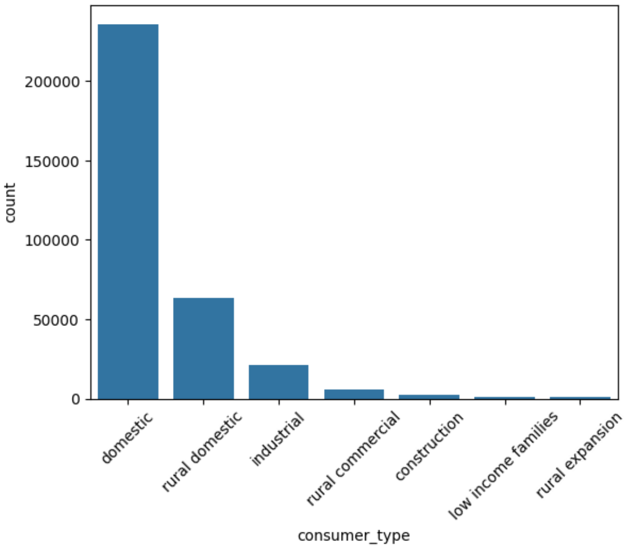


Figure 9: count of consumers per consumer type

The first approach was to use undersampling. We split our model to 'test' and 'train' set and implemented RandomUnderSampler from 'imblearn' module. Initially, the dataset is analyzed to identify the class distribution. If there is a significant disparity between the number of samples belonging to different classes, it is deemed imbalanced. RandomUnderSampler randomly selects a subset of samples from the majority class(es) until the class distribution is balanced. This subset contained the same number of samples as the minority class, effectively reducing the size of the majority class to match the minority class. After the random selection process, a new balanced dataset was formed, where the number of samples belonging to each class was approximately equal. Then we used this newly balanced dataset for training machine learning model. Finally, we liked this model and way of data preparation. As only one class was dominating the dataset, it was natural for us to look for ways of eliminating outliers and reducing the number of inputs for 'domestic' consumer_type.

After that, we followed with our next approach. Our team agreed to test both oversampling and undersampling in the same model to see how they interact with one another. To do this we used SMOTEENN, which performs Synthetic Minority Over-sampling Technique and cleaning using ENN - Edited Nearest Neighbors. It is important to mention, that before applying the algorithm, we split the data into 'train' and 'test' sets using sklearn module. This way, we are sure that the test dataset is unbiased. To justify our choice - we picked SMOTEENN instead of SMOTE because the problem of imbalance is severe and thus unlike traditional undersampling techniques that simply discard majority class samples, SMOTEENN preserves information from the majority class by selectively removing redundant or noisy instances through ENN. This helps retain important characteristics of the majority class while balancing the dataset. Overall, this part of 'imblearn.combine' module offers a comprehensive approach to dealing with class imbalance by leveraging both oversampling and undersampling techniques in a synergistic manner. Following this method allowed us to balance classes for further evaluation and modeling. The size of our original dataset records changed from 329,975 to 443,736 rows. By this method, we added 113,761 new synthetic rows of data, which was later crucial in obtaining reasonable calculations.

Through the process of data preparation, particularly in the context of addressing class imbalance, we learn valuable insights about the dataset itself. This includes understanding the distribution of classes, identifying noisy or irrelevant data points, and recognizing the impact of class imbalance on model performance. SMOTEENN, along with techniques like RandomUnderSampler, provides practical solutions to these challenges, highlighting the importance of thoughtful data preparation in Machine Learning workflows. In summary, SMOTEENN is important for its ability to effectively address class imbalance, preserve information from all classes (majority and minority), reduce overfitting, improve model performance, and provide valuable insights from the data preparation process. By leveraging a combination of oversampling and undersampling techniques, SMOTEENN contributes to the development of more reliable and accurate machine learning models, particularly in tasks with imbalanced class distributions.

2.2 Justification for the opted models

After preprocessing we applied the different primitive machine learning algorithms. These included KNN, decision tree, random forrest, gradient boosting and support vector machine with different hyperparameters. Unfortunately, none of these approaches could achieve a good accuracy, nor a good F1 score. Therefore, we considered two more complex approaches and one very primitive one, which resulted from the understanding of the dataset. These approaches are explained below.

2.2.1 Model 1: Installation Zone Model

The first approach we took was to consider the Installation zone with the consumption and not to consider the year. The month was added to include the seasonal development. What is special about this approach is that we used the uneven data distribution to our advantage. We decided to use two different models. The first model is used to classify the consumer types that are strongly represented and the rest. For this purpose, the training data was divided into four groups. Domestic, industrial, rural domestic and 'grouped_class', where all other label were combined into one group. This allowed us, after balancing, to train with 9,665 entries for each for these four label. If we had worked with all classes and performed a balancing, we would have had only 890 entries for each class.

All entries that have been grouped to the new group are classified in the second model. This is addressed when the first model outputs the label "grouped_class". For the training, we only used data that corresponds to the "grouped_class" and trained the model with its original label. In this case we had to work with the 890 entries per class, but again the model only had to learn a class specification of 4 classes. Since we did not train our models with the consumption of 0, we did not expect a good classification in this case. We decided to exploit the distribution of the classes and classify entries with a consumption of 0 based only on their installation zone. The entries that were in the original installation zone 1-4 were assigned to domestic, all others with a consumption of 0 rural domestic. After setting up the structure, we evaluated different models for each of the two steps. We took into consideration Machine learning algorythms like Support Vector Machine, Decision Tree, Random Forrest, Gradient Descent and K-nearest Neighbor each with different parameters like number of neighbors, learning rate, number of estimators or tree depth.

For each of these approaches we took the overall accuracy to decide which model would be used for our final classification. The best model for classifiing the first four groups was KNN with 9 neighbors with an accuracy of 79 % . The best approach for the second model was the random forrect classifier. With number of estimators of 70 we were able to achieve an accuracy of 51.8 % .

2.2.2 Model 2: Consumption Model

The second model aims to take greater account of consumption. Initially, the approach was to interpret the dataset as a time series. This involved grouping by consumer_number. However, an analysis of the time series per consumer revealed that these are rather incomplete. Consumption is not regularly recorded by every consumer. Given the extent of the data gaps, we refrained from generating missing data by interpolation, as otherwise too much of the data for training would consist of artificial data. Instead, an approach is chosen in which the data is grouped by consumer_number and month, which means that the years are not considered further. For each month, the mean is calculated from all recorded years of a customer. It has already been established in the data understanding that general consumption has not changed significantly over the years. However, this new grouping should take seasonal trends even more into account than the model describited in 2.2.1. This now results in twelve values for consumption.

Unfortunately, it has also been determined here that the twelve months are not suitable as a time series. Missing data is also problematic. Not a single consumer_number has consumption values for all twelve months. The majority only have a value above zero in one to four months. As a result, it does not make sense to fill in missing values by interpolation, because these are not representative and the competition dataset also has such gaps. The training of time series-based models such as Long short-term memory (LSTM) and Dynamic-Time-Warping (DTW) therefore did not appear very promising to us, as the zero values do not reveal any meaningful patterns that can be distinguished on the basis of consumer type. For this reason, the consumption values are also considered as feature values for the second model. This means that the installation zone can also be considered for the training of a model. The same procedure is used for the installation zone as for the first model. Two models are trained with different groupings of the zones.

In order to find out which model is best suited for the preprocessed dataset, the following models are trained with different ones: Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, K-nearest Neighbour and Support Vector Machine. The Decision Tree Classifier delivered the best result with the validation dataset at a maximum depth of 8 with a precision of 93% for domestic and 98% for rural domestic. However, the precision for industrial and the other grouped consumer types is only around 10%. Another decision tree with a maximum depth of 8 is used to classify the remaining consumer types. This achieves an accuracy of 54%. The F1 score for rural commercial is 66%, rural expansion 56%, construction 43% and low income families 27%. When applying the models to the competition dataset, the results of the separate sub-models are combined, with the second sub-model performing the classification for all predicted 'grouped classes' of the first model.

2.2.3 Model 3: Simplest Classifier

Our third approach does not pursue a machine learning approach, but rather follows the idea that a solution to the problem can sometimes be very simple and only an understanding of the initial situation is necessary. This approach follows the characteristics of the extremely undistributed dataset and the separation of the different customer types, that domestic does not occur outside the installation zone 1-4 and is very well represented there, zone 29 and 35 consists only of rural commercial and rural domestic is represented in the other installation zones. We only use a simple case distinction to check in which installations zone this entry is to specify the class.

2.3 Evaluation methodology

The following paragraph describes more detailed how our previously explained models were evaluated and how they performed.

2.3.1 Model 1: Installation Zone Model

For Model 1, the model selection method Two-Way Holdout was chosen, with 80% of the data reserved for training and 20% for evaluation, totaling 65,995 test data points. Due to the complexity associated with using two machine learning models and the tight timeframe, it was decided not to employ the K-Fold cross-validation method to calculate the average results. However, cross-validation is recommended for future work to provide a more comprehensive evaluation. To control for randomness, the 'Random' library was utilized, using the arbitrary value of 42 to ensure consistency in results across each execution. The only modification made to the test data was converting the 'installation zone' field to an integer and creating two new fields, 'Consumer Group' and 'Installation Zone Group', as described in Chapter 2.2.1 . No outliers were removed from the test data to maintain consistency and reliability in obtaining real results. During testing, the first model was used to predict the 'Group Consumer Type'. Based on these first results, only the "grouped_class" was selected for predicting the specific class with the second model. Subsequently, all outlier rows were selected, and their results were replaced with 'domestic' or 'rural domestic' in cases where the 'Installation_zone_group' was 4 or 5. Following these rules, it was possible to calculate the evaluation of Model 1.

	precision	recall	f1-score	support
construction	0.01	0.01	0.01	476
domestic	0.90	0.89	0.90	47,124
industrial	0.16	0.14	0.15	4,244
low income families	0.00	0.01	0.00	192
rural commercial	0.20	0.22	0.21	1,069
rural domestic	0.96	0.68	0.80	12,694
rural expansion	0.01	0.27	0.03	196
accuracy			0.78	65,995
macro avg	0.32	0.32	0.30	65,995
weighted avg	0.84	0.78	0.81	65,995

Table 3: Metrics of Model 1

As seen in table 3, the model achieved an accuracy of 78%. However, the strong point of this model is the identification with greater precision and recall of less common consumer types. The F1 score, which combines precision and recall, serves as the best evaluation criterion for this model. Notably, the F1 scores for ’rural commercial’, ’construction’, ’low income families’, and ’rural expansion’ types are 0.21, 0.01, 0.00, and 0.03, respectively. These scores are low due to limited training data for these groups, but they still indicate good performance compared to other models. Additionally, inaccuracies in identifying ’grouped_class’ in the initial stage contribute to the overall decrease in these metrics.
The confusion matrix of this model is on the figure 10. A small diagonal trend is observed, which is somewhat good.

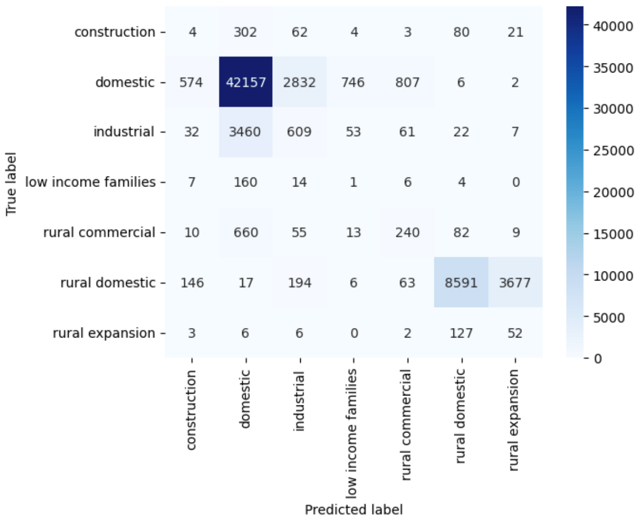


Figure 10: Confusion Matrix of model 1

2.3.2 Model 2: Consumption Model

For the first stage of this model, a similar logic to the previous model was employed. In the initial stage, a grouping of consumer types was performed, where all smaller classes were replaced by ”grouped_class.” The aim of this stage was to identify which grouping each consumer belonged to. In this stage, an accuracy of 52% was achieved, with the best performance observed in identifying the types domestic and rural domestic, with precision ranging between 93% and 98%, as shown in Table 4. However, the precision for identifying the grouped_class was notably low, reaching only 10%.

	precision	recall	f1-score	support
domestic	0.93	0.34	0.50	14,347
grouped_class	0.10	0.29	0.15	758
industrial	0.12	0.74	0.20	1,426
rural domestic	0.98	0.97	0.97	5,351
accuracy			0.52	21,882
macro avg	0.53	0.58	0.45	21,882
weighted avg	0.86	0.52	0.58	21,882

Table 4: Metrics of Model 2 - First Step

The second step is to identify the specific consumer type for those that were identified as grouped class in the first step. In this stage, we found good values as seen in table 5, with an accuracy of 54%, where the F1 score is high for most consumer types, only ”low income families” have a very low value, but this happens due to lack of training data. The F1 scores for less common consumer types like rural commercial, rural expansion, construction, and low-income families ranged from 27% to 66%.

	precision	recall	f1-score	support
construction	0.37	0.51	0.43	196
low income families	0.27	0.28	0.27	61
rural commercial	0.77	0.57	0.66	429
rural expansion	0.46	0.71	0.56	72
accuracy			0.54	758
macro avg	0.47	0.52	0.48	758
weighted avg	0.60	0.54	0.56	758

Table 5: Metrics of Model 2 - Second step

It is important to note that the evaluation with the combination of the two stages was not conducted, but it is crucial to carry out in future works, as this approach allows for determining the final metrics of this model. Considering the low precision of the first step, coupled with the precision of the second stage, it is expected that the final evaluation will also yield low precision.

2.3.3 Model 3: Simplest Classifier

Due to the simplicity of the model, we used the entire dataset for testing because there's no need to separate training data. There's no training involved in this method, only analysis of the Installation Zone, as described in section 2.2.3. We obtained the value as shown in table 6.

	precision	recall	f1-score	support
construction	0.00	0.00	0.00	2,235
domestic	0.90	1.00	0.95	236,167
industrial	0.00	0.00	0.00	21,057
low income families	0.00	0.00	0.00	999
rural commercial	0.77	0.31	0.45	5,541
rural domestic	0.98	0.99	0.98	63,086
rural expansion	0.00	0.00	0.00	890
accuracy			0.91	329,975
macro avg	0.38	0.33	0.34	329,975
weighted avg	0.84	0.91	0.87	329,975

Table 6: Metrics of Model 3

We can notice that for the 'domestic', 'rural commercial', and 'rural domestic' classes, both precision and recall were extremely high, with F1 scores of 0.95, 0.45, and 0.98 respectively. This happens because these types of consumers are the most common in the entire dataset. This model predicts the most common type per installation zone, and the lack of diversification per installation zone leads to very few errors in the model. The accuracy reaches an impressive 91%, but this does not necessarily indicate good performance, as uncommon categories have precision and recall equal to 0.

2.4 Baseline

To compare the results, we utilized the dummy algorithm as a baseline. This model predicts the most common consumer type, which in this case is 'domestic.' Employing K-Fold cross-validation, we obtained an accuracy of 71.57% and an F1 score of 11.92%.

Due to the imbalance in consumer types within the dataset, it presents a challenging environment for model training. When compared to the baseline, only Model 3 achieved a higher accuracy, reaching 91%. However, when considering the F1 score, higher values were obtained in Models 1 and 3, with 30% and 34% respectively. Unfortunately, final metrics for Model 2 were not available for comparison.

Based on the F1 metric, Models 1 and 3 demonstrate effectiveness as they outperform the baseline. Model 3 emerges as the superior among the trained models, boasting the highest overall accuracy and F1 score. However, Model 1 excels in identifying less common consumer types, albeit performing less satisfactorily when considering the dataset as a whole.

The comparison underscores the trade-offs between different models: while Model 3 achieves high accuracy across the board, Model 1 exhibits strength in capturing nuances within less common consumer types. These findings suggest that the choice of model depends on the specific goals and priorities of the analysis. Additionally, further investigation into Model 2's final metrics would provide a more comprehensive assessment of its performance and its potential contribution to the overall model ensemble.

3 Results and Discussion

After classifying the competition dataset, the results were provided by the lecturers and are as follows:

	accuracy	macro_avg_precision	macro_avg_recall	macro_avg_f1
Model 1: Installation Zone Model	0.5904	0.3674	0.3290	0.3118
Model 2: Consumption Model	0.5678	0.3780	0.3788	0.3666
Model 3: Simplest Classifier	0.6807	0.3242	0.3459	0.3191

Table 7: Final Results

According to our evaluation, the first model has a macro average recall of 32 % and was therefore rated around 4 % poorer than in the competition dataset.Unfortunately, we do not have an overall evaluation for the second model and cannot say exactly whether the performance in the competition dataset was well predicted with 37.88 % . For the third model, we predicted a macro average recall of 33 %, which was very close to the competition with a value of 32.42 %. Based on the comparison with other groups, we were surprised that we created the second best model according to the macro average recall.

3.1 Model selection and evaluation

Due to time constraints, we have primarily focused on the approaches described in chapters 2.2.1 and 2.2.2. Within these, we took the time to try out different algorithms and do a little fine tuning. Since we only had two models in addition to our very primitive one, the situation of selecting our best three approaches did not arise.

3.2 Description and interpretation of the results

The results of the trained models are not yet perfect. Summarising the consumption in the second model only had a minor influence on the Installation_zone feature. It should be analysed in more detail how a better estimation can be made based on the consumption. It should also be investigated how time series can still be used in a practical way. If there was more time, we would have mixed the combination of different approaches and tried to find a better combination. In addition, one aspect that would probably have given us a better result, but would have taken too much time, would have been optimizing the grouping of the installation zones. We also thought about whether it would be better to apply one classification model to each of the installation zones. Either with the grouping or without the grouping, the model for each zone could focus on a smaller selection of labels. Unfortunately, our team lacked the capacity to pursue LSTM or other more complex approaches.

4 Conclusion

Looking back on the project, it wasn't a lack of ideas but primarily a lack of time. There were a few other approaches that we actually wanted to try out, but which were not dealt with due to our priorities. Everyone in our team was involved in data understanding. In Braganca, we split up due to time constraints. Dominik primarily worked on balancing and considered which approach has which advantages and disadvantages. Gustavo developed a method to compare our models against each other in a representative way. Andre and Malte primarily dealt with the training of the models and the associated pre-processing