Prediction of the Number of Days for Sale in the Real Estate Asset Market

Antonio Moratilla[1], Eugenio Fernández[1],Ángel Álvarez[1], Álvaro F. Narciso[2]
[1](*Computer Science and Artificial Intelligence Department, Alcalá University, Spain)*
[2](*CitizenLab Project)*

ABSTRACT: *This article aims to correctly predict real estate TOM (Time on Market) in Madrid based on areal dataset, after doing some data cleaning and exploration. In particular, the article is evaluating different ML techniques to make predictions: LASO, RIDGE, KNNR, XGBR, LGBM. Clarifying comments have been made throughout the report and there is also a conclusion at the end. The article shows the results achieved in a research project with real data.*

KEYWORDS - *Machine Learning, Prediction, Real Estate Assets Market, TOM Time On Market.*

I. INTRODUCTION

The real estate market in Spain has always been a significant pillar of the national economy. In recent years, this market has experienced various tensions, impacting both sales and rental segments.

The sales sector of Spanish real estate has seen fluctuating trends, primarily influenced by economic shifts and policy changes. Post the 2008 global financial crisis, Spain's property market underwent a substantial correction, leading to a significant drop in property prices. However, the market showed resilience, and by the mid-2010s, it began to recover, driven mainly by domestic demand. The government has been grappling with balancing the need for foreign investment with the housing needs of its citizens.

The rental market in Spain has been under even more strain. In urban centers, especially in Madrid and Barcelona, rental prices have skyrocketed in recent years. This surge is partly attributed to the popularity of short-term rental platforms like Airbnb, which have reduced the supply of long-term rental properties, so finding affordable rental housing in these cities has become increasingly challenging for locals. The Spanish government has taken steps to regulate the short-term rental market and protect the rights of long-term residents. These measures change how real state properties are commercialized, so it's important to be able to predict how a property is going to make on market.

There are a lot of factors that must be taken in consideration to be able to predict how fast a property is going to be sell: the property itself, the surroundings (schools nearby, sport zones…), near services, etc. Analyzing all these factors, Nikiforu et al (2002) [1] research on how DOP (degree of overpricing), TOM (time on market) and SP (selling price) are related. There are previous works by Knight [2] and Glower et al [3] analyzing marketing strategies and the market reaction to a new property selling point, while concluding the main strategy to assign a selling price: the property' maintenance costs. That SP relates tom TOM on the margin offered to market.

Arrazola [4] analyzes the housing market on Spain on a 34-year series, trying to estimate the market's behavior, centering its efforts on SP and TOM given various stocks levels. On Caldera and Johansson [5] works arises how housing are far sensible on USA real estate market than other countries, using data from Riddel [6] and Ball et al [2010]. That behavior shows how housing markets depends on different variables, and not all of them have equally importance on different markets. Steiner [6] analyze Swiss market, and Kenny [7] makes also on Irish market, arising different elasticity on demand curves.

Also, the elasticity of income varies on different countries, altering how variables are taken in consideration, as Malpezzi and Maclennan [8] shows.

Other macroeconomic indicators have its own effect on house market, like interest rates as shown on Kenny [9] and Steiner [6].

As has been shown, being able to predict how much time it's going to take to sell or loan a property on real estate market has multiple factor that leads to a complex problem. In order to extract knowledge on these kind of complex problems, Machine Learning (ML) techniques has been used, combined with data mining and statistics. There are publications centered on TOM of a property like Hengshu Zhu [10], where Linear Regression (LR), Lasso and Decision Trees (DT) are used. Other works like Ermolin [11] makes use of DT to predict TOM on a 7-day window.

II. METHODOLOGY

The work methodology followed is integrated into the project's work methodology that supports the work carried out. This methodology is defined based on domains, processes, and tasks, and aligns with corporate Data Science workflows.

The methodology of this article covers the domains of data processing, feature engineering, and model development.

In the data processing section, tasks such as data collection, data adequacy, statistical selection of data, establishment of sets, data cleaning, analysis of data balance, and exploratory causal and data analysis are addressed.

Regarding feature engineering, work is carried out on feature selection, feature extraction through dimensional reduction, and feature transformation.

Finally, in terms of the model development domain, tasks primarily involve model evaluation, selection, and tuning.

As a final element within the methodology, the technical characteristics of the work environment used will be briefly discussed.

III. DATA DESCRIPTION AND ANALYSIS

The data used is from Idealista dataset for 2020 year on Madrid data. This is market' real data of 116282 rows, with 53 variables in its original form.

Work based on previous methodology arises a 25,4% of duplicated rows, with 43 numeric variables and 10 categorical variables.

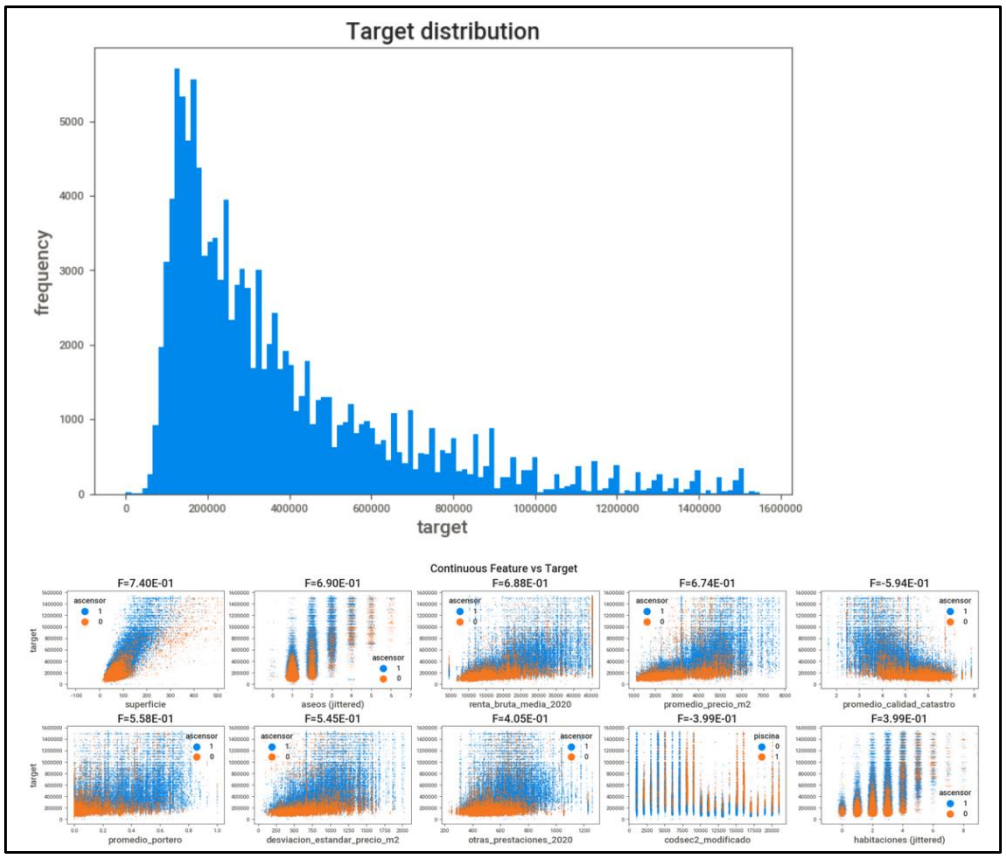

Fig. 1: Target distribution for features.

As shown on Fig 1, features are evenly distributed taken on consideration its values. The features from data and its datatypes are listed on Fig. 2:

superficie	float64
habitaciones	int64
aseos	int64
barrio	int64
localizacion_piso	int64
ascensor	int64
tiene_aparcamiento	int64
trastero	int64
piscina	int64
jardin	int64
terraza	int64
calefaccion	int64
aire_acondicionado	int64
dias_en_venta	int64
codsec2_modificado	int64
edad_media_2020	float64
poblacion_2020	int64
hogares_unipersonales_2020	float64
porcentaje_espannoles_2020	float64
p_mayores_64_2020	float64
p_menores_18_2020	float64
personas_por_hogar_2020	int64
media_renta_por_unidad_consumo_2020	int64
mediana_renta_por_unidad_consumo_2020	int64
renta_bruta_media_por_hogar_2020	int64
renta_bruta_media_por_persona_2020	int64
renta_neta_media_por_hogar_2020	int64
renta_neta_media_por_persona_2020	int64
renta_bruta_media_2020	int64
salario_2020	int64
desempleo_2020	int64
pensiones_2020	int64
otras_prestaciones_2020	int64
otros_ingresos_2020	int64
indice_gini_2020	float64
renta_p80/p20_2020	float64
promedio_anno_catastro	float64
promedio_precio_m2	float64
desviacion_estandar_precio_m2	float64
promedio_altura_edificio	float64
promedio_atico	float64
promedio_calidad_catastro	float64
promedio_vivienda_nueva	float64
promedio_reformado	float64
promedio_distancia_centro	float64
promedio_distancia_eje	float64
promedio_estado	float64
promedio_estudio	float64
promedio_portero	float64
promedio_piso	float64
promedio_viviendas_edificio	float64
promedio_precio	float64
target	int64

Fig. 2: Features and datatypes used on dataset.

Further individualized analysis shows different behavior over the data distribution of each feature as shown on Fig 3.

To select significant features over the whole dataset, a correlation analysis is made using the Pearson method, where a value of 1 implies correlation between 2 features, a value of 0 means no correlation, and a value of -1 implies negative correlation. The results of this analysis are shown on Fig. 4 and Fig. 5.

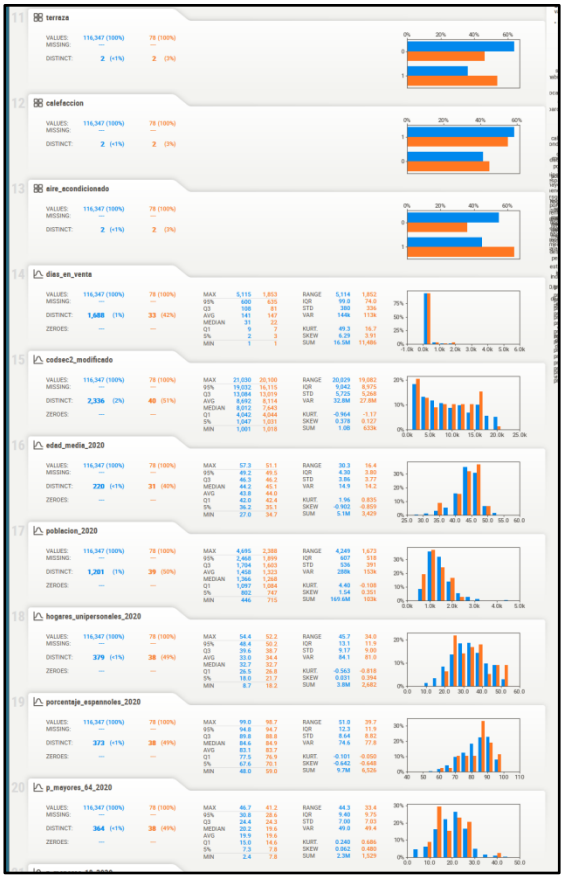

Fig. 3: Exploratory Analysis.

As stated on Fig 5 and Fig. 6, there's a high correlation between features of the dataset, implying relations between them at business level. Those relations can be broken on some categories, as economical category as there's data about income of the zone where the asset is located: all assets inside that geographical zone (a block, a neighborhood…) will share its features values, so internal correlation is expected.

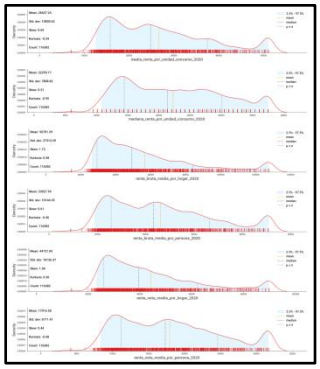

Fig. 4: Detail of economical category features.

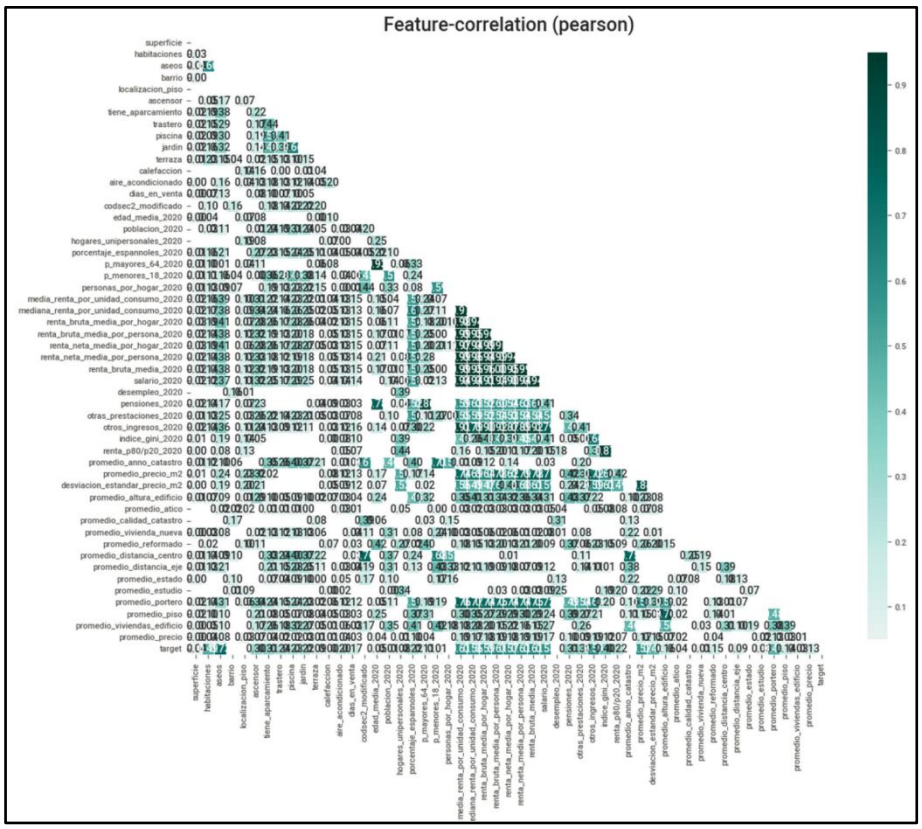

Fig 5. Positive correlation

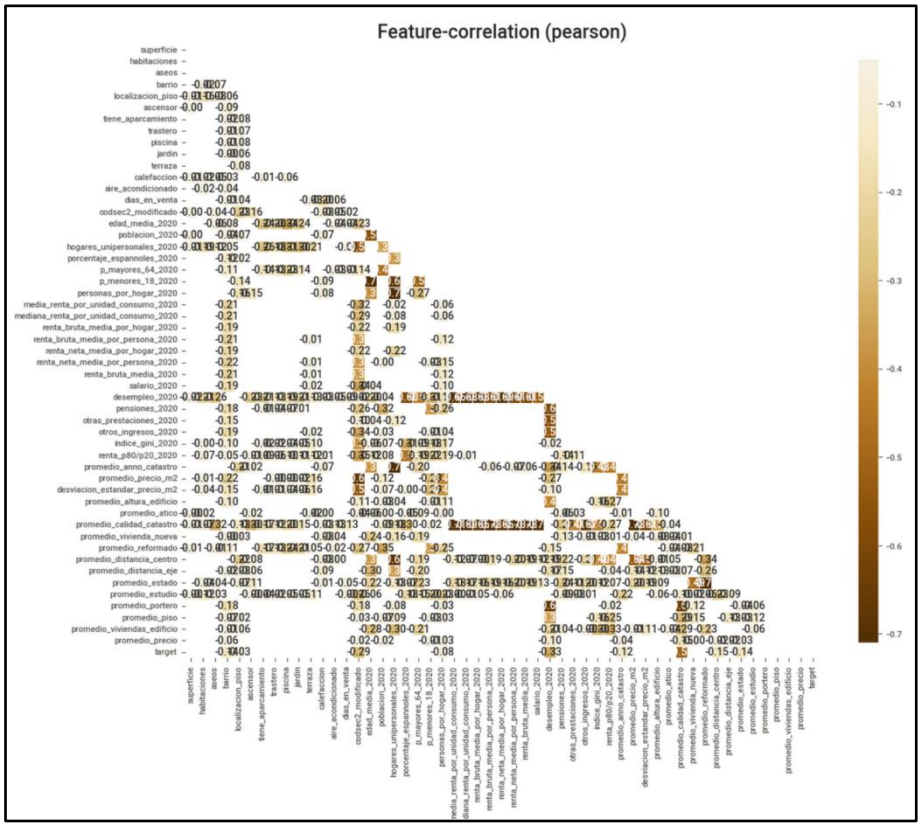

Fig 6. Negative correlation

IV. MODEL SELECTION

To predict the TOM of an asset, the feature "dias_en_venta" (days on sell) is used. As stated in reference section, ML algorithms has been

used to predict various features over a dataset. We split the original dataset on two main categories: sell and loan. Each category is subsequently split on type of asset (home, house, garage or commercial), and split again in training and test dataset.

Those datasets are feed to the 5 candidates' methods for prediction: LASSO, RIDGE, KNNR, XGBR, and LGBM in order to create a model for each one and test against the appropriate data subset. These methods are based on 3 different kinds of algorithms: linear regressions, clustering and decision trees.

LASSO and RIDGE methods are algorithm to reduce the problem complexity by modeling relationship between a dependent variable (which may be a vector) and one or more explanatory variables, fitting regularized least squares model. They're usually used for feature selection and overfitting prevention. They're linear regression-based algorithms. LASSO tends to shrink coefficients to zero, reducing the complexity, while RIDGE don't and is used when almost all features are relevant, especially in multicollinearity problems.

KNNR (K-Nearest Neighbors Regressor) is based on nearest neighbor clustering techniques. It works based on proximity of previous samples, and its K parameter is critical to boost precision of its predictions. It's sensitive to data scaling and irrelevant dimensions.

XGBR (XGBoostRegresor) and LGBM (Light Gradient Boosting Machine) are based on Decision Trees algorithms, using a gradient boost method to optimize its results. Both are fast techniques, being LGBM a refined version with good performance on categorical data. XGBR is popular in ML academic research, being both efficient and having techniques to prevent overfitting.

Added to the base models, we create a Voting system where all models ads-up different results in order to test a mixture of experts.

V. Deployment Details

All training and evaluation model has been made using programs written in Python language, using the appropriate libraries where possible, or designing and implementing our own algorithms where needed. Those programs are designed to run on a web environment, like a Jupyterserver, using a Voilà server.

The deployment of the solutions has been made using containers. Containers is a technology that makes possible to create a solution on a computer and send them to other computers or a cluster of them seamlessly, without problems with dependencies or incompatible versions. It allows to be run on cloud resources with minimal modifications. We've created PowerBI visualization for data exploration and selection. A dedicate server with Intel XEON processor, 32 Gb of RAM and 2Tb disk on RAID1 configuration over an Ubuntu operating system has been used.

VI. Results

Results obtained are calculated for each data subset, and combined on the next Table 1:

Modelo	Training_Score	Test_Score	R2
LASO	0.977	0.942	0.942
RIDGE	0.970	0.938	0.938
KNNR	0.891	0.866	0.866
XGBR	0.813	0.721	0.721
LGBM	0.699	0.701	0.701
VOTING	0.699	0.701	0.701

Table 1. Results.

On Figure 7 are shown the detail of actual TOM and the predicted TOM.

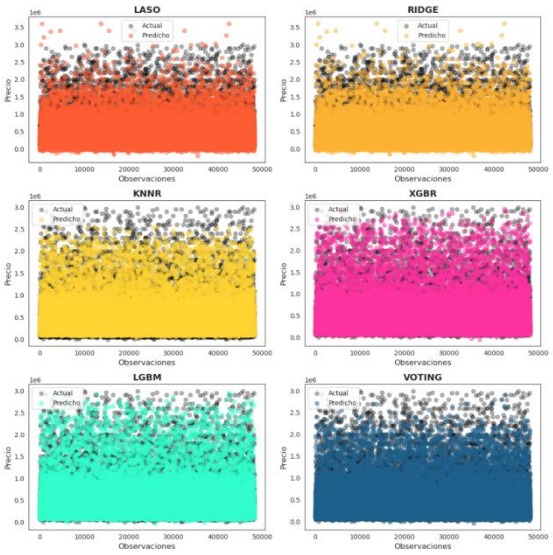

Fig. 7: Result details

On Fig. 8 the user interface is shown while Madrid data is being visualized. At the right side of the figure, there's a dispersion graph of data points for TOM.

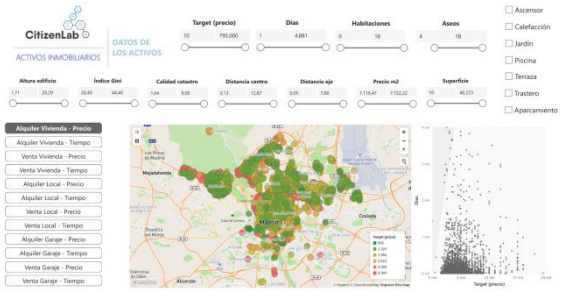

Fig 8: User Interface

The results on Table 1 confirms the initial analysis done in the preprocessing phase of the methodology, where correlations over the features on the dataset where found, and some linearity aroused. Given the 5 prediction methods used, both linear methods, LASSO and RIDGE, shows better performance, and between the two, LASSO is slightly better. This is coherent with findings on preprocessing phase as LASSO tries to eliminate features simplifying the result model.

With Clustering KNNR scoring near linear methods, the Tree based methods XGBR and LGBM falls behind on performance values, albeit being used in other studies.

VII. CONCLUSION

This study uses real estate data from Madrid area (Spain) to evaluate 5 methods to predict TOD of assets in the market. The preprocessing phase of the study shows some characteristics that may influence the performance of the methods evaluated.

The results shown are consistent with the findings, and linear regression techniques give the best performance on this scenario.

The main conclusion on the article is that the best method to predict TOM on a real state dataset cannot be stated for all cases, as it depends on the inner information structure of the dataset. Preprocessing phase on data is strictly necessary to guide the election of the candidate prediction' methods. While previous works center their effort on choosing some technique over others, our work shows this cannot be generalized in this kind of problem for all datasets, as they may vary its internal information structure. Future work may be done to characterize the datasets to be able of classify them and guide the prediction model' selection.

Acknowledgements

This study was conducted as part of the CitizenLab project, financed by the Community of Madrid.

REFERENCES

[1] PhilipposNikiforou, Thomas Dimopoulos and Petros Sivitanides (2022) Identifying how the time on the market affects the selling price: a case study of residential properties in Paphos (Cyprus) urban area. Journal of European Real Estate Research Vol. 15 No. 3, 2022 pp. 368-386.

[2] Knight, J.R. (2002), "Listing price, time on market, and ultimate selling price: causes and effects of listing price changes", Real Estate Economics, Vol. 30 No. 2, pp. 213-237.

[3] Glower, M., Haurin, D.R. and Hendershott, P.H. (1998), "Selling time and selling price: the influence of seller motivation", Real Estate Economics, Vol. 26 No. 4, pp. 719-740.

[4] Arrazola, M., de Hevia, J. Romero, D. & Sanz-Sanz, J.F. (2014) Determinants of the Spanish housing over three decades and three booms: long run supply and demand elasticities. Working Paper 13/2014. September 2014. Working Papers in Public Finance. Victoria. University of Wellington.

[5] Caldera Sánchez, A y A. Johansson (2011) The Price responsiveness of housing supply in OECD countries, OECD Economics Department Working Papers, 837, OECD.

[6] Steiner, E. (2010) Estimating a stock-flow model for the Swiss housing market. Swiss Society of Economics and Statistics, 146, No. 3, pp 601 – 627.

[7] Kenny, G. (1999), Modelling the demand and supply sides of the housing market: Evidence from Ireland. Economic Modelling, 16(3), 389-409.

[8] Malpezzi, S. and D. Maclennan (2001), "The long-run price elasticity of supply of new residential construction in the United States and the United Kingdom". Journal of Housing Economics, 10, pp. 278-306.

[9] Kenny, J (2003) Asymmetric adjustment cost and the dynamics of housing supply,

Economic Modelling, 20, pp. 1097 – 1111.

[10] H. X. Hengshu Zhu, "Days on market: measuring liquidity in real estate markets," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 393–402, Beijing, China, 2016.

[11] S. V. Ermolin, Predicting Days-on-Market for Residential Real Estate Sales, Department of Computer ScienceStanford University, Stanford, CA, USA, 2016, http://cs229.stanford.edu/proj2016/report/ermolin_predicting_Days_on_market_for_Residential_Real_Estate_Sales_report.pdf.

[12] V. Fonti, Research Paper in Business Analytics: Feature Selection with LASSO, VU Amsterdam, Amsterdam, Netherlands, 2017.