Applying Machine Learning to real estate data

Organisation:	Statistics Poland
Author(s):	Dominik Dabrowski, Marta Kruczek-Szepel, Klaudia Peszat, Krystyna Piatkowska
Date:	16.12.2021
Version:	1.0

Spis treści

Background

The Statistics Poland is Poland's chief government executive agency charged with collecting and publishing statistics related to the country's economy, population and society, at the national and local levels. In previous ML Project Marta Kruczek-Szepel and Krystyna Piatkowska worked on applying Machine learning methods on ECOICOP data[1]. The project had satisfying results and it was decided to try those techniques on different scopes of data.

The information scope of the real estate market statistics produced by the official statistics is relatively limited to the price, size, the number of rooms and the basic amenities of the premises. While these data provide some basic information, purchasing decisions are influenced by many additional factors, which are not a subject of observation by registers or statistical surveys. These factors are related to a widely understood standard of living, including e.g. the availability of parking spaces, terraces, gardens, ensuring the safety of the place through security services, etc.

This project aims to explore new data source in order to prepare system for monitoring the real estate market on an ongoing basis by analysing publicly available online data sources (websites of real estate agents, search engines for real estate sales offers presenting ads from sales and rent market). The analysis assumes preparing an infrastructure and methodological note for producing on an at least monthly basis experimental statistics on the evolution of observed variables (including characteristic of real estate and the prices) on the lowest possible territorial level and possibly beyond the standard administrative divisions. The analysis will gather the basic real estate market data, such as the surface, number of rooms and the price of the real estate, but they will also attempt to cover additional qualitative information, such as type of kitchen and security to measure the standard of the available real estate. This task can be completed by implementing machine learning techniques to classify the unstructured information hidden in the text description of the offer.

[1] https://statswiki.unece.org/download/attachments/285216428/ML_WP1_CC_Poland.pdf

Data

Input data

Except the structured basic information (price, area) presented on the real estate offers websites, multiple information are inside the object description. Unlike the data processed in the ECOICOP project, here was a need to process long text variables storing information on many different subjects without no structure.

At the beginning of the project we assessed web data sources, based on the number of advertisements, number of variables in advertisement, response level of the site, structure of HTML code and a possible access via API. We also checked how the sources present the information, if they use captcha, provide filtering, listing the pages, using dynamic loading of offers, have up to date content. This process has been carried out on 10 portals, with half of them assessed positively and 2 definitely rejected (due to lack of labelled information or major problems with page operations).

Besides the basic variables like price, area, number of rooms we searched through the portals for 32 different variables referring to the standard of the apartment. On the base of commonness we decided to eventually focus on two – Kitchen type and Security.

Table 1. Variables considered for the project.

Variable name	Variable description
balcony	if the object has balcony, terrace or loggia
heat_type	type of heating system in the object
elev	if the building has elevator
park_space	if the object has a parking space
fur	if the object is furnished
condition	if the object is renovated/to be renovated/luxury
const_type	material used for construction
new_dev	new development or not
owner	form of ownership (private property, cooperative)
deposit	cost paid with the beginnig of the agreement (offers for rent)
fees	value of maintenance fees
two_floor	if the object is two-staged
tv	if the object has TV sets
refr	if the object has refrigerator
wash_mach	if the object has washing machine
oven	if the object has Oven
dishwasher	if the object has dishwasher
dryer	if the object has Dryer
bathtub	if the object has Bathtub/ jacuzzi
air_cond	if the object has air conditioning
internet	if the object has internet
security_type	if there is any of the security systems
garage	if the object has a garage
base_type	if the object has one of the following: storage room, cellar or basement

kitchen_type	type of the kitchen
garden	if the object has a garden
pets	if pets are accepted (only rent)
students	if students are accepted (only rent)
win	window type

As the project was carried out for learning purposes we were not obliged by any regulations in case of choosing target variables. The ones we chosed were: Kitchen – whether the object has opened kitchen or separate; Security – wheter the object provides any security features e.g. concierge, monitoring system, fenced area, entry phone.

Following the assessment we decided to choose one, the most convenient portal for further work – domoferty.pl and we respected both the portals regulations regarding the intellectual law rules and robots.txt file.

In the first step, we manually copied about 100 descriptions to test if the ML methods are better than random classification. The results turned out to be promising, so we started to prepare the database structure to collect a larger set of data.

The data was finally obtained through the offered API of domoferty.pl. Text and numeric data extracted made up the input data. The data was partially structured. We did not wanted to highly interrupt the portal, so we downloaded the data by night and focused on possibly more specific group of offers. Finally we decided to collect only data regarding offers of apartments for sale which is in Poland a bigger market than the rents.

On the collecting step the data was cleaned from multiple white spaces, newlines and other disrupting signs. We did not processed any further cleaning. The dataset is very raw with possible duplicated offers (if they were added by different agencies), also containing bailout auctions (which did not have information on the variables we were interested in) or offers from not yet constructed objects where kitchens were sometime not yet decided to be opened or separate. This could possibly made the set overcoveraged.

All the programming processes were carried out in Python and SQL database. The offers were read through an API and stored in a list. Each list was then send to the table on the SQL server. The API connection allowed us to download 12105 offers from all sixteen country regions. The amount of offers by territorial level are presented in the Table 10.

The database not only stores scraped data, but also helps to manage text classification stages (a dedicated table was created to governance the work on manual classification). We have extracted from the API the following variables:

- Advertisement id
- Advertisement description
- Price
- Currency
- Area in m2
- Address (town, street)
- Kitch (kitchen)
- Secure (security)
- Kitch_gr (assignment kitchen type to numeric values)
- Secure_gr (assignment security type to numeric values)

- Descr (description of the offer)
- Ad_added (date of adding the offer)
- Ad_updated (date of updating the offer)
- Terr_id (id of territorial level, voivodship)
- Trans_id (transaction Id)
- Source_id
- Comment_id

The database also included various additional variables for the project purposes:

- Insert_date (date of collecting the data through API)
- Modify_date (if the offer was updated in our databes)
- Inserted_by (code of the user that manually classified the data through a dedicated app)
- Modified_by (code of user who was the last one classifying the offer)
- Kitch_type (manually classified on the basis of Descr)
- Secure_type (manually classified on the basis of Descr)

Next step was to analyze descriptions and available basic variables (transaction type, price, surface, description, address, kitchen type, security features), from the content of HTML pages. The dataset contains a wide variety of property descriptions. From short, not containing much information, to very extensive, presenting a large amount of information, but also a lot of redundant text not referring to the content of the presented property.

Beneath are presented the statistics of the length of the variable ‘description’.

Table 2. Five shortest lengths of variable ‘description’ by number of signs.

Length	Number of words
94	7
110	14
128	16
134	14
151	18

Table 3. Five longest lengths of variable ‘description’ by number of signs.

Length	Number of words
14141	1995
12923	1787
12842	1779
9797	1312
9797	1312

Some of the offers were presenting multiple apartments for sale. An example of one of those cases is presented below. It offers 14 separate apartments in one building. All the flats differs by the area size. In case of those offers we still did not have a solution for their non-manual extraction from the dataset. One of the problem is lack of patterns to use in searching such descriptions as the the expression 'm2' may also refer to subsequent rooms.

Example of an offer presenting multiple apartments for sale:

Wielofunkcyjne lokale i mieszkania - idealna inwestycja przy ul. św. Franciszka Salezego, ŚródmieścieMieszkania znajdują się w 10 piętrowym bloku na 1 piętrze.
Całkowita powierzchnia: 372 m2
W skład wchodzą:
1. kawalerka 19.12 m2
2. kawalerka 18.34 m2
3. kawalerka 18.51 m2
4. kawalerka 19.18 m2
5. kawalerka 19.91 m2
6. kawalerka 19.78 m2
7. kawalerka 19.19 m2
8. mieszkanie 2 pokojowe 34.59 m2
9. mieszkanie 2 pokojowe 37.05 m2
10. kawalerka 19.66 m2
11. kawalerka 18.15 m2
12. kawalerka 18.59 m2
13. mieszkanie 2 pokojowe 25.17 m2
14. kawalerka z balkonem 20.5 m2
Oraz wspólny korytarz o pow. 64 m2.
Okolica:Do Centrum 5-10 minut, dobrze skomunikowane miejsce ze wszystkimi dzielnicami Warszawy. Bezpośrednie połączenia z Centrum tramwajowe i autobusowe. W sąsiedztwie znajduje się rzeka i nadwiślańskie bulwary. W pobliżu: sąsiedztwo sklepów, obiektów sportowych i nowoczesnych biurowców, bardzo blisko komunikacja miejska, apteka, poczta, centrum kultury. Powyższa oferta ma charakter informacyjny i nie stanowi oferty handlowej w rozumieniu art. 66 §1 Kodeksu Cywilnego.Podana ulica nieruchomości przy ogłoszeniu ma na celu wskazania przybliżonej lokalizacji z dużą dokładnością.Bezpłatnie pomagamy znaleźć najkorzystniejszą na rynku ofertę kredytową.Na życzenie w ramach prezentacji nieruchomości możliwość przeprowadzenia przez firmę budowlaną oceny stanu mieszkania i wycenę remontu.

For the report purposes the data was additionally cleaned from outlayers. The results are aggregated in tables 13-16. For the project purposes and machine learning algorythms we used a raw dataset. This was due to the fact of lack of time for appropriate data cleaning. The next step of extended testing of machine learning methods conducted in the next year are planned to consider higher level of data cleaning and input data preparation.

Data preparation

Unfortunately for the project and report purposes it was impossible to translate the descriptions into English. The fact that this variable is unstructured and often full of language mistakes makes it unable to automatically translate the full scope, maintaining a satisfactory level of translation quality.

For preparing the data to manual classification the duplicates according variable ‘description’ were excluded from the scope. We excluded descriptions that were duplicates after cleaning the data from space, blank lines, numbers, and special signs. It allowed us to treat as a duplicate all offers that were referring to same standard apartments within one property. This was crucial to exclude repeating offers on the process of manual classification. Below are presented descriptions of three offers from one building differing only (in descriptions) by apartment numbers.

Table 4. Sample of duplicated descriptions referring to different offers.

Description[2]	Price	Currency	Area	Address
KUPNO OD DEWELOPERA, 0% prowizji, PLANOWANE TERMINY ZAKOŃCZENIA INWESTYCJI - budynek A - IV kwartał 2021, budynek B - II kwartał 2022Budynek A lokal: 1.A.8.13, balkon [Mieszkania z opcją SMART HOME, pakiet antysmogowy]Cena brutto m. p. w garażu podziemnym: - 35 000 zł (naziemne - 18 000 zł) - ILOŚĆ OGRANICZONA !!	351241	PLN	34.52	Hetmańska, Poznań
KUPNO OD DEWELOPERA, 0% prowizji, PLANOWANE TERMINY ZAKOŃCZENIA INWESTYCJI - budynek A - IV kwartał 2021, budynek B - II kwartał 2022Budynek A lokal: 1.A.8.11, balkon [Mieszkania z opcją SMART HOME, pakiet antysmogowy]Cena brutto m. p. w garażu podziemnym: - 35 000 zł (naziemne - 18 000 zł) - ILOŚĆ OGRANICZONA !!	447526	PLN	48.91	Hetmańska, Poznań
KUPNO OD DEWELOPERA, 0% prowizji, PLANOWANE TERMINY ZAKOŃCZENIA INWESTYCJI - budynek A - IV kwartał 2021, budynek B - II kwartał 2022Budynek A lokal: 1.A.8.6, balkon [Mieszkania z opcją SMART HOME, pakiet antysmogowy]Cena brutto m. p. w garażu podziemnym: - 35 000 zł (naziemne - 18 000 zł) - ILOŚĆ OGRANICZONA !!	459450	PLN	51.05	Hetmańska, Poznań

The manual classification was prepared by a team of 4 people. They managed to manually classify almost 6,7 thous. of ads. from 8.8 thous. of unique ads.

For the process of manual classification there was a dedicated application prepared.

It was created with the use of python Flask library. The principle of operation was to present the first not yet classified offer from an assigned voivodship. It colours the keywords (ochrona – *security*, kuchnia – *kitchen*) that may be used to choose the category which we are interested in (for example if the kitchen is separate or not, if there is some kind of security system).

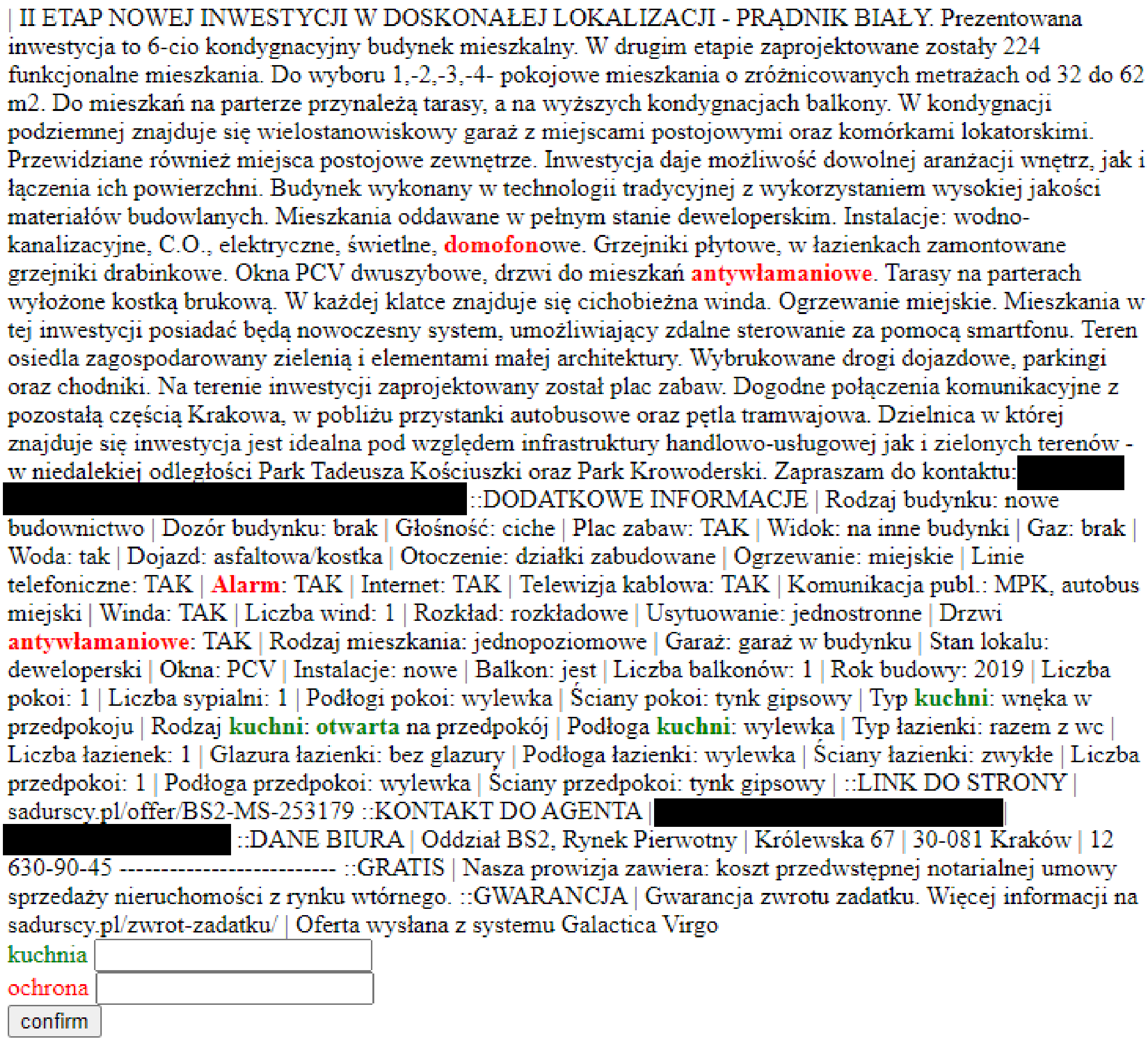
In case of security information we decided only to assess if the security systems exists (any of them) or not (if there was no information). While analysing the kitchen information we wanted to assess if the kitchen is opened or separate and when no information occurred we marked the offer as ‘no information’.

The list of keywords was chosen on the base of the experience from working on the first manually copied base of 100 offers. There were two lists prepared. One for searching information on kitchen types, and the other on the information on security type. The lists of words contained 10 expressions for searching kitchen information and 19 for security information.

[2] Descriptions in the table were cut, so they could fit into the table.

All of the four manually classifying people were working on different set of data, because the application each time presented only the offers which were not yet classified. The workers were free to choose the voivodship they worked on. If there was a problem in deciding on specified information the worker could mark the record as 'for consultation'. Which was later assessed by the team.

Below is presented a view of one of the offers with highlighted expressions:



| II ETAP NOWEJ INWESTYCJI W DOSKONAŁEJ LOKALIZACJI - PRĄDNIK BIAŁY. Prezentowana inwestycja to 6-cio kondygnacyjny budynek mieszkalny. W drugim etapie zaprojektowane zostały 224 funkcjonalne mieszkania. Do wyboru 1,-2,-3,-4- pokojowe mieszkania o zróżnicowanych metrażach od 32 do 62 m2. Do mieszkań na parterze przynależą tarasy, a na wyższych kondygnacjach balkony. W kondygnacji podziemnej znajduje się wielostanowiskowy garaż z miejscami postojowymi oraz komórkami lokatorskimi. Przewidziane również miejsca postojowe zewnętrze. Inwestycja daje możliwość dowolnej aranżacji wnętrz, jak i łączenia ich powierzchni. Budynek wykonany w technologii tradycyjnej z wykorzystaniem wysokiej jakości materiałów budowlanych. Mieszkania oddawane w pełnym stanie deweloperskim. Instalacje: wodno-kanalizacyjne, C.O., elektryczne, świetlne, **domofon**owe. Grzejniki płytowe, w łazienkach zamontowane grzejniki drabinkowe. Okna PCV dwuszybowe, drzwi do mieszkań **antywłamaniowe**. Tarasy na parterach wyłożone kostką brukową. W każdej klatce znajduje się cichobieżna winda. Ogrzewanie miejskie. Mieszkania w tej inwestycji posiadać będą nowoczesny system, umożliwiający zdalne sterowanie za pomocą smartfonu. Teren osiedla zagospodarowany zielenią i elementami małej architektury. Wybrukowane drogi dojazdowe, parkingi oraz chodniki. Na terenie inwestycji zaprojektowany został plac zabaw. Dogodne połączenia komunikacyjne z pozostałą częścią Krakowa, w pobliżu przystanki autobusowe oraz pętla tramwajowa. Dzielnica w której znajduje się inwestycja jest idealna pod względem infrastruktury handlowo-usługowej jak i zielonych terenów - w niedalekiej odległości Park Tadeusza Kościuszki oraz Park Krowoderski. Zapraszam do kontaktu: [redacted] ::DODATKOWE INFORMACJE | Rodzaj budynku: nowe budownictwo | Dozór budynku: brak | Głośność: ciche | Plac zabaw: TAK | Widok: na inne budynki | Gaz: brak | Woda: tak | Dojazd: asfaltowa/kostka | Otoczenie: działki zabudowane | Ogrzewanie: miejskie | Linie telefoniczne: TAK | **Alarm**: TAK | Internet: TAK | Telewizja kablowa: TAK | Komunikacja publ.: MPK, autobus miejski | Winda: TAK | Liczba wind: 1 | Rozkład: rozkładowe | Usytuowanie: jednostronne | Drzwi **antywłamaniowe**: TAK | Rodzaj mieszkania: jednopoziomowe | Garaż: garaż w budynku | Stan lokalu: deweloperski | Okna: PCV | Instalacje: nowe | Balkon: jest | Liczba balkonów: 1 | Rok budowy: 2019 | Liczba pokoi: 1 | Liczba sypialni: 1 | Podłogi pokoi: wylewka | Ściany pokoi: tynk gipsowy | Typ **kuchni**: wnęka w przedpokoju | Rodzaj **kuchni**: **otwarta** na przedpokój | Podłoga **kuchni**: wylewka | Typ łazienki: razem z wc | Liczba łazienek: 1 | Glazura łazienki: bez glazury | Podłoga łazienki: wylewka | Ściany łazienki: zwykłe | Liczba przedpokoi: 1 | Podłoga przedpokoi: wylewka | Ściany przedpokoi: tynk gipsowy | ::LINK DO STRONY | sadurscy.pl/offer/BS2-MS-253179 ::KONTAKT DO AGENTA | [redacted] ::DANE BIURA | Oddział BS2, Rynek Pierwotny | Królewska 67 | 30-081 Kraków | 12 630-90-45 ------------------------- ::GRATIS | Nasza prowizja zawiera: koszt przedwstępnej notarialnej umowy sprzedaży nieruchomości z rynku wtórnego. ::GWARANCJA | Gwarancja zwrotu zadatku. Więcej informacji na sadurscy.pl/zwrot-zadatku/ | Oferta wysłana z systemu Galactica Virgo

The application is each time saving results to dedicated database after confirming the entered expressions. There was no possibility to return to previously input data.

Machine learning solution

While Scikit-Learn Python library is being used, we apply vectorizer (CountVectorizer) in order to prepare our dataset for Machine Learning classification. It converts text data into vectors of numbers (because ML models can process only numerical data). By using CountVectorizer we divide each text on parts divided by blank spaces and allocates the amount of appearances of each word within the record.

Additionally, some normalization is conducted by the vectorizer as well. All the words are lowercase, the punctuation and numbers are removed. We have not provided any extra normalization so far. We do not deal with any abbreviations or acronyms and we do not convert numeric expressions and verbal-numeric expressions to verbal form.

Finally we assigned to a testing sample a set of 500 offers, for the validation sample a set of next 500 offers. The rest 5627 offers was the training set.

Models tried

For the project purposes we used logistic regression model.
At the beginning the set was prepared by data conversion using CountVectorizer which converted text data into vectors.

Used parameters:
C: [0.1, 1, 2, 3]
fit_intercept: tested both
Class_weight: balanced
Solver: ["newton-cg", "lbfgs", "liblinear", "sag", "saga"]
multi_class: { ‘ovr’, ‘multinomial’}
max_iter = 400

Table 5. Four best parameters specifications for logistic regression model for variable kitchen_type:

C	fit_intercept	class_weight	solver	multi_class	no_iterations	validate_accuracy	train_accuracy	harmonic_mean
2.0	False	None (=1)	newton-cg	multinomial	400	0,924	0,998578283	0,924
2.0	False	None (=1)	lbfgs	multinomial	400	0,924	0,998578283	0,924
3.0	False	None (=1)	newton-cg	multinomial	400	0,922	0,998933712	0,922
3.0	False	None (=1)	lbfgs	multinomial	400	0,922	0,998933712	0,922

Table 6. Four best parameters specifications for logistic regression model for variable secure_type:

C	fit_intercept	class_weight	solver	multi_class	no_iterations	validate_accuracy	train_accuracy	harmonic_mean
3.0	False	balanced	newton-cg	multinomial	400	0,926	0,998222	0,926
3.0	False	balanced	newton-cg	ovr	400	0,93	0,997867	0,93
2.0	False	None (=1)	newton-cg	multinomial	400	0,918	0,999467	0,918
3.0	False	None (=1)	newton-cg	multinomial	400	0,918	0,999467	0,918

Almost all set of tried parameters gave results with accuracy level higher than 90%. Some of the tried parameters sets were not successfully finalized within the number of given iterations.

First trials with Logistic Regression have been conducted during the project. Other methods, such as Naive Bayes, Random Forest and SVM will be tested in the next step as well as verification of possibilities to use neural networks for classification. We would also like to use the transformers methods that are getting their popularity in recent years.

Model(s) finally selected and quality criteria used (e.g. accuracy, time)

The parameters of logistic regression selected for both variables, after fine-tuning and testing all the possibilities:
C: 3
fit_intercept: False
Class_weight: None
Solver: newton-cg
multi_class: multinomial
max_iter = 400

Results

We wanted our classifier based on a machine learning algorithm to predict the category for every input data given. In fact we were searching for the category with the highest probability for each offer. We saved the best results for the testing set of data to the file where we can compare the correct category with the predicted one. We are also saving the probability of this category.

In the table below there is a list of automatically assigned categories of kitchen type with the probability of the assignment. The results on the test set looks promising as we have more than 72% of the data with probability over 95%. Still in this group we tracked 6% of observations falsly allocated. In the group of observations with probability lower than 95% we observed a high share (60%) of falsely assigned kitchen categories.

We believe that it is possible to use the probability results to create a threshold for choosing the offers for manual classification.

The result table (below) presents that even with a 100% probability there is a mistake that the algorithm did. In the example no. 14 the program falsely assigned kitchen type as ‘separate’. This may be caused because of high complexity of the descriptions and not fully prepared set of manually classified offers. Nevertheless the statistical models are not truly infallible.

Table 7. Predicted categories with highest probability in the test set (500 offers).

No.	description[3]	kitchen_type	autocode	probability
1	| NOWE OSIEDLE MIESZKANIOWE W	aneks	aneks	1
2	| MISTRZEJOWICE - NOWE MIESZKA	aneks	aneks	1
3	KUPUJĄCY 0% PROWIZJI! AP 19 Pr	aneks	aneks	1
4	3-pok. mieszkanie z balkonem n	odrębna	odrębna	1
5	Biuro nieruchomości Pierwsze P	aneks	aneks	1
6	Do sprzedaży bardzo ładne prze	aneks	aneks	1

[3] Descriptions in the table were cut to the length of 30, so they could fit into the table.

7	| NOWE OSIEDLE MIESZKANIOWE W	aneks	aneks	1
8	| NOWA INWESTYCJA MIESZKANIOWA	aneks	aneks	1
9	| NOWA INWESTYCJA POŁOŻONA NA	aneks	aneks	1
10	Syndyk masy upadłości J.J. – o	brak informacji	brak informacji	1
11	| WOLA DUCHACKA - NOWOCZESNY K	aneks	aneks	1
12	| Oferta dla wymagającego klie	aneks	aneks	1
13	| PRĄDNIK BIAŁY - NOWY ETAP IN	aneks	aneks	1
14	| Do sprzedaży 2-pokojowe mies	aneks	odrębna	1
15	Zapraszam do obejrzenia i zaku	aneks	aneks	1
16	NA SPRZEDAŻ MIESZKANIE 2-POKOJ	aneks	aneks	1
17	Witam serdecznie, Mam przyjemn	aneks	aneks	1
18	| Oferujemy na sprzedaż atrakc	odrębna	odrębna	1
19	KUPNO OD INWESTORA, 0% prowizj	brak informacji	brak informacji	1
20	NA SPRZEDAŻ MIESZKANIE 2-POKOJ	aneks	aneks	1
21	Nowy kompleks budynków wieloro	brak informacji	brak informacji	1
22	| II ETAP NOWEJ INWESTYCJI W D	aneks	aneks	1
23	KUPNO OD INWESTORA, 0% prowizj	brak informacji	brak informacji	1
24	W ofercie biura Expander Nieru	aneks	aneks	1
25	| WOLA DUCHACKA - KAMERALNA IN	aneks	aneks	1
26	SAWKA Nieruchomości oferuje lo	aneks	aneks	1
27	NA SPRZEDAŻ MIESZKANIE 2-POKOJ	aneks	aneks	1
28	Polecam do sprzedaży rewelacyj	aneks	aneks	1
29	NA SPRZEDAŻ MIESZKANIE 2-POKOJ	aneks	aneks	1
30	NA SPRZEDAŻ MIESZKANIE 2-POKOJ	aneks	aneks	1
31	NA SPRZEDAŻ MIESZKANIE 2-POKOJ	aneks	aneks	1
32	Na sprzedaż przestronne 4 poko	aneks	aneks	1

In the tables below are presented classification reports of the applied model with the precision, recall and F1-score for each category of each data set. The support is the number of offers in this category in the given set.

Table 8. The classification report for logistic regression of variable kitchen_type

	Category	**Precision**	**Recall**	**f1-score**	**Support**
TRAINING DATA	aneks	1.00	1.00	1.00	2437
	brak_informacji	1.00	1.00	1.00	1803
	odrębna	1.00	1.00	1.00	1387
VALIDATION DATA	aneks	0.94	0.96	0.95	217
	brak_informacji	0.91	0.90	0.90	154
	odrębna	0.91	0.89	0.90	129
TEST DATA	aneks	0.93	0.93	0.93	227
	brak_informacji	0.86	0.84	0.85	165
	odrębna	0.78	0.81	0.79	108

Table 9. The classification report for logistic regression of variable security_type

	Category	Precision	Recall	f1-score	Support
TRAINING DATA	brak_informacji	1.00	1.00	1.00	3797
	tak	1.00	1.00	1.00	1828
VALIDATION DATA	brak_informacji	0.91	0.97	0.94	323
	tak	0.94	0.82	0.88	177
TEST DATA	brak_informacji	0.91	0.97	0.94	331
	tak	0.93	0.82	0.87	169

Next step was outputting the results of classification for the whole scope (11915 observations). They are presented in Table 11 and Table 12.

Code/programming language

Python 3.8
Scikit-Learn library

Conclusions and lessons learned

In the on-line real estate advertisements the offers tend to have very long descriptions. It still could be a problem in teaching algorithms to classify them correctly. Especially when there is a high share of offers with almost identical descriptions, but referring to different object.

In fact manual classification of couple thousand of observations, especially with such an extended variables is also a challenge. We need to find solutions that will help us to deal with this problem, taking into consideration that there are still more information that may be extracted from the data, e.g. mentioned in Table 1.

In some cases we had to deal with not precised information of variables in the apartments descriptions. Especially in new constructions the information of kitchen type was not still precised in the direction we were interested.

It was easier to manually classify variable ‘security’, as we decided to classify it only on two categories (Yes – if any of security features exists or No information – if there was nothing mentioned in the offered). In the case of kitchen it was harder to decide if the kitchen is separate or opened. Some offers had not precised descriptions, like “kitchen is semi-opened”.

Another problem was caused by offers presenting multiple apartments with different specifications. They also should be extracted out of the scope, but it would need dedicated methods to at first try to find such offers and then delete them from dataset used or divide them into separate ones.

During the data processing and preparing the dataset, especially the descriptions of the offers, we should have used cleaned version of the descriptions for the machine learning algorithms. Otherwise we give the full descriptions with phone numbers, prices etc. without giving the model an information on their importance level.

We believe that the manual classification could have been done better, especially when we would classify the observations parallelly by two people. But still the achieved results are satisfying. The results achieved allow us to believe that the undertaken further actions will lead to the creation of an automatic classification tool with a high probability of correctness of the results. We are aware that there are still lots of methodological work to prepare, especially to check the representativeness of the data sources and widening the scope of data.

Still most of the problems occurred as a result of not precisely prepared methodology. However the project was especially undertaken for learning purposes. The team which worked on previous machine learning project was widened to share knowledge within the organisation.

Preparing the project have led us to create a dedicated infrastructure for data acquisition within the official statistics network. After consulting the IT department it was decided to set a server with access to dedicated disk space. That allowed us to work semi-isolated from the internal network and have better performance executing scripts. This solution will be used in future data science projects targeting bigdata acquired form internet data sources.

Collaboration with other statistical organisations, universities, etc

There has been no collaboration with organizations outside the Statistics Poland structures.

Summary and next steps

The project is a proof of concept for testing a new data source which could supplement the real estate statistics in Poland. But if the project results are not satisfactory, they will not be implemented yet in the production.

As the subject is new in Statistics Poland, and we do not have a fully operational tools, we do not implement results into official statistics.

Such statistics on real estate data (acquired from the internet data sources) were never published, the work undertaken in the ML project is aimed at extending knowledge and skills.

We have not developed any unified methodology to carry out the classification and certainly the manual classification made, may be burdened with human error. But nevertheless the achieved results showed that manual classification provided satisfactory results.

We are going to develop the prepared tools for further testing but also we consider using neural networks methods or the transformers methods. This still needs lots of methodological and on-desk work as probably there will be a need of preparing two versions of the dataset as in the case of neural networks we would need to use the descriptions of real estate untouched. Neural network tokenizers the order of words in the sentence, the punctuation and numbers are essential to understand the language and the context.

In the next phase we would like to create more preprocess steps and try different machine learning models.

There may be also an attempt made to test the algorithms behaviour when stop-words libraries would be introduced or used different set of seed for random generated values.

List of tables

Table 10. Number of observations by voivodships.

Voivodship	Number of offers	Number of offers without duplicates
DOLNOŚLĄSKIE	866	708
KUJAWSKO-POMORSKIE	426	387
LUBELSKIE	41	38
LUBUSKIE	217	148
ŁÓDZKIE	77	52
MAŁOPOLSKIE	4334	2290
MAZOWIECKIE	3332	2934
OPOLSKIE	44	44
PODKARPACKIE	265	254
PODLASKIE	110	95
POMORSKIE	208	193
ŚLĄSKIE	477	372
ŚWIĘTOKRZYSKIE	414	354
WARMIŃSKO-MAZURSKIE	64	64
WIELKOPOLSKIE	683	541
ZACHODNIOPOMORSKIE	547	407
Overall	12105	8881

Table 11. Share of kitchen type and number of offers by territory.

Voivodship	Opened	No information	Separate	Number of offers
DOLNOŚLĄSKIE	29,8%	35,0%	35,2%	856
KUJAWSKO-POMORSKIE	21,4%	64,2%	14,4%	425
LUBELSKIE	27,5%	52,5%	20,0%	40
LUBUSKIE	21,0%	49,5%	29,4%	214
ŁÓDZKIE	40,0%	50,7%	9,3%	75
MAŁOPOLSKIE	72,9%	16,0%	11,2%	4248
MAZOWIECKIE	38,3%	28,0%	33,7%	3315
OPOLSKIE	18,6%	41,9%	39,5%	43
PODKARPACKIE	59,5%	33,0%	7,6%	264
PODLASKIE	48,1%	36,1%	15,7%	108
POMORSKIE	55,8%	33,2%	11,1%	199
ŚLĄSKIE	31,7%	58,7%	9,5%	441
ŚWIĘTOKRZYSKIE	60,4%	3,6%	35,9%	412
WARMIŃSKO-MAZURSKIE	33,9%	53,6%	12,5%	56
WIELKOPOLSKIE	33,7%	38,9%	27,4%	676
ZACHODNIOPOMORSKIE	45,3%	45,9%	8,8%	543
OVERALL	50,4%	28,3%	21,3%	11915

Table 12. Share of security equipment occurrence and number of offers by territory.

Voivodship	brak informacji	tak	Number of Offers
DOLNOŚLĄSKIE	68,3%	31,7%	856
KUJAWSKO-POMORSKIE	68,2%	31,8%	425
LUBELSKIE	100,0%	0,0%	40
LUBUSKIE	69,2%	30,8%	214
ŁÓDZKIE	57,3%	42,7%	75
MAŁOPOLSKIE	38,7%	61,3%	4248
MAZOWIECKIE	75,8%	24,2%	3315
OPOLSKIE	83,7%	16,3%	43
PODKARPACKIE	84,1%	15,9%	264
PODLASKIE	85,2%	14,8%	108
POMORSKIE	57,8%	42,2%	199
ŚLĄSKIE	69,2%	30,8%	441
ŚWIĘTOKRZYSKIE	70,6%	29,4%	412
WARMIŃSKO-MAZURSKIE	78,6%	21,4%	56
WIELKOPOLSKIE	67,2%	32,8%	676
ZACHODNIOPOMORSKIE	69,1%	30,9%	543
OVERALL	**60,4%**	**39,6%**	11915

Table 13. Statistics for offered price in PLN by voivodships

Voivodship	count	mean	std	min	25%	50%	75%	max
DOLNOŚLĄSKIE	856.0	488544.6	362825.6	52000.0	303000.0	426074.5	565250.0	4899991.0
KUJAWSKO-POMORSKIE	425.0	343249.8	137875.5	43000.0	260000.0	320000.0	384900.0	1400000.0
LUBELSKIE	40.0	393767.0	123582.9	235000.0	316250.0	377500.0	403000.0	802000.0
LUBUSKIE	214.0	299236.0	102883.9	50000.0	242199.9	283881.1	345527.4	825000.0
ŁÓDZKIE	75.0	357858.1	120385.8	96600.0	294500.0	370000.0	434961.0	700000.0
MAŁOPOLSKIE	4248.0	608381.1	365990.6	115000.0	425362.0	521544.0	687018.5	5653500.0
MAZOWIECKIE	3315.0	759793.6	559948.0	83000.0	499000.0	630257.0	820000.0	10161310.0
OPOLSKIE	43.0	281093.0	165920.6	29000.0	184000.0	250000.0	340000.0	759000.0
PODKARPACKIE	264.0	424420.5	175094.5	130000.0	328785.0	390000.0	475000.0	1690065.0
PODLASKIE	108.0	402006.7	149968.5	109000.0	308000.0	359000.0	453475.0	1100000.0
POMORSKIE	199.0	642519.4	504786.5	209000.0	379000.0	515000.0	777821.5	5900000.0
ŚLĄSKIE	441.0	353669.5	161393.2	79000.0	249000.0	317000.0	416540.0	1500000.0
ŚWIĘTOKRZYSKIE	412.0	408178.3	173454.1	55000.0	280000.0	377924.0	500000.0	1161185.0
WARMIŃSKO-MAZURSKIE	56.0	404560.3	187338.7	125000.0	270930.0	372000.0	470000.0	1007950.0
WIELKOPOLSKIE	676.0	411444.4	179033.3	95000.0	293576.5	384960.0	492261.2	1818571.0
ZACHODNIOPOMORSKIE	543.0	553061.4	375825.2	141350.0	360396.7	480931.5	618076.5	3525747.0

Table 14. Statistics for area of the offered apartments in m2 by voivodships

Voivodship	count	mean	std	min	25%	50%	75%	max
DOLNOŚLĄSKIE	856.0	58.6	28.1	16.1	42.0	53.5	67.3	305.8
KUJAWSKO-POMORSKIE	425.0	53.7	18.8	18.4	40.8	50.9	63.4	186.0
LUBELSKIE	40.0	52.5	15.5	20.6	40.6	55.2	60.6	89.4
LUBUSKIE	214.0	60.0	18.4	13.6	48.3	55.6	67.1	140.0
ŁÓDZKIE	75.0	53.2	19.9	13.1	40.4	50.6	62.4	130.0
MAŁOPOLSKIE	4248.0	55.6	26.1	10.0	39.1	51.2	65.6	260.0
MAZOWIECKIE	3315.0	59.2	28.9	15.5	41.4	53.0	67.9	350.4
OPOLSKIE	43.0	75.6	48.0	33.0	50.4	64.0	81.5	325.0
PODKARPACKIE	264.0	61.0	18.4	20.1	48.6	58.2	70.9	132.0
PODLASKIE	108.0	53.7	18.1	24.2	42.0	49.2	60.7	140.0
POMORSKIE	199.0	61.0	24.0	25.5	47.1	55.5	69.6	196.0
ŚLĄSKIE	441.0	57.9	23.5	24.0	43.1	53.8	66.6	252.1
ŚWIĘTOKRZYSKIE	412.0	61.3	21.5	22.9	46.8	59.0	70.9	136.6
WARMIŃSKO-MAZURSKIE	56.0	63.5	20.4	35.2	48.3	59.6	76.6	113.2
WIELKOPOLSKIE	676.0	54.9	24.9	16.8	39.7	52.0	65.4	229.4
ZACHODNIOPOMORSKIE	543.0	53.3	29.9	23.9	30.7	43.1	69.4	221.1

Table 15. Statistics for date of add of the advertisement in days by voivodships (0 = today)

Voivodship	count	mean	std	min	25%	50%	75%	max
DOLNOŚLĄSKIE	856.0	10.4	12.1	0.0	1.0	6.0	13.0	59.0
KUJAWSKO-POMORSKIE	425.0	9.7	12.9	0.0	1.0	5.0	13.0	57.0
LUBELSKIE	40.0	18.0	18.0	0.0	3.0	11.5	23.0	58.0
LUBUSKIE	214.0	21.7	13.9	0.0	15.0	20.0	24.0	59.0
ŁÓDZKIE	75.0	12.1	11.6	1.0	6.0	6.0	15.0	55.0
MAŁOPOLSKIE	4248.0	8.3	10.1	0.0	5.0	5.0	5.0	58.0
MAZOWIECKIE	3315.0	8.0	12.1	0.0	1.0	1.0	11.0	60.0
OPOLSKIE	43.0	17.4	13.1	1.0	9.0	14.0	22.0	51.0
PODKARPACKIE	264.0	20.0	14.6	0.0	6.0	17.0	34.0	59.0
PODLASKIE	108.0	10.2	12.3	0.0	1.0	8.0	15.0	57.0
POMORSKIE	199.0	12.0	12.9	0.0	1.0	6.0	15.0	59.0
ŚLĄSKIE	441.0	15.6	13.8	0.0	3.0	13.0	26.0	59.0
ŚWIĘTOKRZYSKIE	412.0	19.3	11.4	0.0	14.0	20.0	27.0	44.0
WARMIŃSKO-MAZURSKIE	56.0	6.5	12.2	0.0	1.0	1.0	3.0	56.0
WIELKOPOLSKIE	676.0	13.4	11.3	0.0	5.0	11.0	21.0	57.0
ZACHODNIOPOMORSKIE	543.0	22.4	13.0	0.0	10.0	28.0	28.0	59.0

Table 16. Statistics for price per square meter of the offered apartments in PLN

Voivodship	count	mean	std	min	25%	50%	75%	max
DOLNOŚLĄSKIE	856.0	8643.6	3356.1	990.5	5898.9	8788.0	10763.2	27403.3
KUJAWSKO-POMORSKIE	425.0	6478.5	1417.3	1954.5	5595.2	6333.3	7307.7	12334.8
LUBELSKIE	40.0	7725.3	1633.9	4000.0	6512.2	7347.3	8689.9	12542.4
LUBUSKIE	214.0	5089.0	1251.1	641.5	4477.7	4990.0	5800.0	8254.6
ŁÓDZKIE	75.0	6948.8	1631.5	2650.0	6133.4	7200.0	8000.0	9500.0
MAŁOPOLSKIE	4248.0	11279.4	3571.2	2464.4	9000.0	10707.0	12570.9	40000.0
MAZOWIECKIE	3315.0	12899.4	4388.8	4266.7	10225.4	12383.1	14865.0	55233.2
OPOLSKIE	43.0	4101.2	2051.4	537.0	2339.4	4630.3	5465.4	9000.2
PODKARPACKIE	264.0	7047.5	1652.6	2000.0	6331.6	6900.0	7850.5	13894.4
PODLASKIE	108.0	7680.6	1895.7	1524.4	6245.9	7750.9	8984.3	11734.7
POMORSKIE	199.0	11068.1	8946.1	3544.3	7155.1	8964.0	11597.3	109259.3
ŚLĄSKIE	441.0	6246.3	1859.2	2042.8	5053.5	5948.4	7250.0	13043.5
ŚWIĘTOKRZYSKIE	412.0	6654.1	1364.9	1672.7	5855.8	6858.6	7700.0	10626.2
WARMIŃSKO-MAZURSKIE	56.0	6351.8	1762.3	1875.0	5481.5	6053.5	7276.1	10750.0
WIELKOPOLSKIE	676.0	8209.1	3462.1	1669.6	5715.0	7844.5	9262.5	17656.0
ZACHODNIOPOMORSKIE	543.0	11500.4	5676.6	2400.2	6990.8	9690.0	15600.0	53258.3