IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 15, 2021, accepted April 30, 2021, date of publication May 5, 2021, date of current version May 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3077860

Can Web Search Queries Predict Prices Change on the Real Estate Market?

NINA RIZUN AND ANNA BAJ-ROGOWSKA
Department of Informatics in Management, Gdańsk University of Technology, 80-233 Gdańsk, Poland
Corresponding author: Nina Rizun (nina.rizun@pg.edu.pl)

ABSTRACT This study aims to explore whether the intensity of internet searches, according to the Google Trends search volume index (SVI), is a predictor of changes in real estate prices. The motivation of this study is the possibility to extend the understanding of the extra predictive power of Google search engine query volume of future housing price change (shift direction) by (i) the introduction of a research approach that combines the advantages of the complementary use of cross-correlation analysis and machine learning classification algorithms; (ii) applying the multi-class HPI values classifier which allows predicting the housing price increase, decrease or relative stability; (iii) exploiting the SVI that relates to interests in both 'real estate' and 'credit to buy real estate'; (iv) evaluation of the introduced approach in the context of the Polish real estate market. The main theoretical contribution of our work is a confirmation that the freely available information regarding Google user searches can provide an in-depth insight into enriching the generally accepted statistics on supply and demand in the real estate market. From the practical perspective, this research confirms that SVI can be associated as a sole determinant to anticipate the housing price change with time-lag sufficient for making decisions regarding the purchase (sale) of individual property or the real estate market control. Such findings can be also helpful for researchers who intend to use Google Trends data as an extra variable from demand side to improve the prediction accuracy if it is included in the model which is based on the existing housing prices determinants.

INDEX TERMS Big data utilization, cross-correlation analysis, machine learning classification, Google trends (GT), house price index (HPI), prediction, real estate market, search volume index (SVI), time-lag.

I. INTRODUCTION

The analysis of data from search engines is an interesting area of research in big data literature. Google search engine is an excellent platform for observing people's activities related to searching for information in various fields. It offers an instant response that reflects the needs and interests of its users. Search engine traffic seen as the volume of search requests submitted by users to search engines on the web can be used to track and, in some cases, to anticipate the dynamics of social phenomena [4]. Authors term this occurrence as a sort of wisdom of the crowd effect. According to this theory [9], a large group of diverse people will produce better and more solid forecasts than even the most qualified decision-maker. As practice shows, the mean of the group's guesses is more precise than any individual's estimation. The idea of crowd wisdom is not that a group will always give us the right answer, but will indicate a better answer than any individual could provide. The findings of [10] showed that experience diversity, participant independence, and network decentralization are all positively related to crowd performance. Diversity means each member of the crowd has some private information and a distinct interpretation. Independence stands for individuals' opinions that are not determined by those around them. Decentralization means that crowd members have their own specializations and can learn from their own knowledge sources.

The value of aggregated search data as a predictive tool is noted by academics and practitioners in various fields. The first to notice these particular features in scientific research were Ginsberg *et al.* [3]. Their findings suggested that tracking queries in Google search engine gives the possibility of indicating the population in which flu is prevalent. Ultimately, denoting SVI as a very efficient indicator of the spread of the flu virus, led to the development of an epidemic tracking tool called Google FluTrends. In turn, the results

The associate editor coordinating the review of this manuscript and approving it for publication was Justin Zhang.

of [5] provided evidence that SVI might improve the unemployment rate (in the example of Romania) and should be considered in view of providing better forecasts to support government decisions. Bordino *et al.* [4] proved that daily trading volumes of stocks traded in NASDAQ-100 are correlated with daily volumes of queries related to the same stocks and, more precisely, query volumes in many cases anticipate the peaks of trading by one day or more. Takeda and Wakao [6] examined the relationship between online search intensity and stock-trading behaviour in the Japanese market. They found a correlation between the value of the searched company names (Nikkei 225 index) and the volume of their turnover. Moreover, Tao *et al.* [7] created an SVI driven model based on Google Trends data to forecast crude oil prices.

Search engine data, which are mainly perceived as a representative index of human attention, have been widely used in predicting different phenomena. A study conducted by the National Association of Realtors [8] reveals that the share of home buyers who used the Internet to search for a home increased to an all-time high of 97%. Thus, it could be said that the real estate purchase process starts with a search engine. This user approach generates a lot of valuable information for researchers in discovering the usefulness of these data for the housing market. There are a number of studies concerning the usage of a search volume index (SVI) for prediction purposes (e.g. house prices and transaction or sale volume, etc.) on various real estate markets, for example, on the English housing market [39], [12], [35], the US market [34], [37], [2], and on the Indian market [1]. However, most of these studies are based on hedonic-based regression models to build predictive models and study the predictive power of Google Trends data only as an additional enhancement of the models which are based on traditional determinants of housing price accuracy [12], [39], [34], [35]. At the same time, there are more and more promising research results regarding the application of machine learning models to investigate the possibility of using the potential of search engine data in different areas of economic activity [25], [43]–[45].

Thus, the motivation behind the main *goal* of this study is the possibility to extend the understanding of the Google user search volume extra *predictive power* of future housing price change by (i) the introduction of a complex research approach that combines the advantages of the complementary use of cross-correlation analysis and machine learning classification algorithms; (ii) applying the multi-class HPI values classifier which allows predicting the housing price increase, decrease or relative stability; (iii) exploiting the Google user search volume that relates to interests in both 'real estate' and 'credit to buy real estate'; (iv) evaluation of the introduced approach in the context of the Polish real estate market.

The central *research question* (RQ) is: What research potential does the Google search volume have related to users' interests in both 'real estate' and 'credit to buy real estate', as an sole determinant to predict the changes (shift direction) in real estate price? We divided the main research question into two sub-questions, namely: (RQ1) Is there a correlation between instances of changes in the Google user queries search volume (SVI) and the price activity (HPI) in the Polish real estate market; if yes, are these instances synchronous or characterized by a particular time-lag? (RQ2) Whether changes in the users' Google search activity (SVI) with respect to information about 'real estate' market offers and 'credit conditions to buy real estate' have an extra predictive power of future housing price changes, in addition to all existing determinants of housing price; if yes - is machine learning classification good enough for predicting price shift direction in the Polish housing market based only on Google search activity?

The *methods* used to answer the questions include (1) expert and correlation analysis approach to adjust a target sample with the aim to improve the accuracy of the predictive model; (2) a cross-correlation analysis to confirm the assumption of the presence of a correlation between the instances of changes in the user search volume and prices in the real estate market with a particular time-lag; (3) machine learning classification algorithms to explore the predictive potential of the Google user search volume (expressed by SVI) in the context of HPI in the Polish market.

The main *theoretical* contribution of this paper is (1) a demonstration of the *powerful potential* of analytical methods for studying user activity in the Internet space to expand the capabilities of price predictive models; and (2) a confirmation that the freely available information regarding Google user searches can provide an in-depth insight into *enriching the generally accepted statistics* on supply and demand in the real estate market. The main *practical* contribution of our research is that Google user search volume can be considered as an sole determinant to anticipate the housing price shift direction with time-lag sufficient for making decisions regarding the purchase (sale) of individual property or the real estate market control.

The rest of the paper is organized as follows. The next section reviews the relevant literature. Section 3 describes the data used in the analysis and the methodology employed. In Section 4, we present our findings. Section 5 discusses our results. We conclude in Section 6 and briefly delineate future directions for our research.

II. LITERATURE REVIEW

A. RESEARCH WITH GOOGLE SEARCH VOLUME DATA

Search query data were not publicly available until the launch of Google Trends in 2006. Two years later, Google Insights was launched as an advanced and more detailed service that provided data to users on search trends. The earliest and the most significant studies based on using GT data, such as [3] and [11], have demonstrated that Google search data can help forecast the spread of the flu and the unemployment rate, respectively. Moreover, [12] found that Google Trends is useful for forecasting many economic activities, e.g. car sales, home sales, retail, and travel. They revealed that if there is an

TABLE 1. Cluster extraction.

No	Name of cluster	Article keywords	Number of papers
1	**CANCER**	e.g. melanoma, prostate and testicular cancers, colorectal cancer, breast tumour, oral cancer, skin cancer, lung cancer, kidney cancer, …	199 articles on the topic 'cancer'
2	**EPIDEMIOLOGY**	e.g. Chikungunya virus, influenza, a digital epidemiological study in connection with the disciplines of dentistry, Hepatitis D virus (HDV) infection, allergy, …	150 articles on the topic 'epidemiology'
3	**CORONAVIRUS**	e.g. coronavirus, covid-19, sars, mers, …	97 articles on the topic 'coronavirus', 23 articles - 'sars', 33 'mers' and 142 - 'covid-19'
4	**TOURISM**	e.g. tourism demand, tourism forecasting, tracking the changing trends in ecotourism over a decade, research on comparing and ranking state government tourism websites in the Indian context, econometric models for improvement in the prediction of tourist arrivals, …	108 articles on the topic 'tourism'
5	**SUICIDE**	e.g. suicide prediction, suicide prevention; mental health, celebrity suicide, …	97 articles on the topic 'suicide'
6	**BITCOIN; CRYPTOCURRENCY**	e.g. models on the bitcoin transaction network for forecasting the bitcoin price movement, models for showing which country has the biggest demand on particular cryptocurrencies, the growth rate of Google Trends shows statistically significant effects on Bitcoin returns volatility, there is a relevant degree of correlation of Google Trends and Tweet volume data with the price of Bitcoin, and with no significant relation with the sentiments of tweets. …	63 articles on the topic 'bitcoin' and 9 -'cryptocurrency'
7	**STOCK MARKET**	e.g. stock prediction, stock market anomaly, market efficiency, volatility forecasting, …	38 articles on the topic 'stock market'

increase in searches for ''real estate property'' in a particular location, it is possible to predict that housing sales will soon increase in this location. These papers are some of the most influential papers among the studies based on Google Trends.

Nowadays, Google Trends is a popular target for exploring search engine data because it is easy to access, is provided free of charge, and has the potential to explain and predict better the different indicators for which official data are provided with a delay. This has been reflected in the number of indexed papers in the Scopus database. Fig. 1 shows the number of studies from 1st January 2006 to 19th September 2020 which are dedicated to the research of GT data.

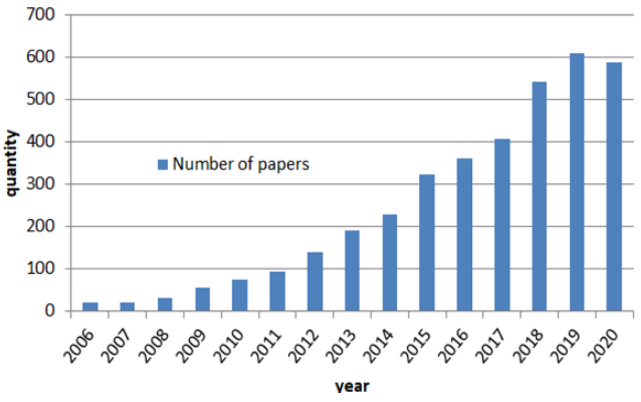

FIGURE 1. Publication trends of literature related to google trends (until September 2020).

B. RESEARCH BASED ON GOOGLE TRENDS DATA USED FOR FORECASTING

The number of papers dedicated to the topic of Google Trends is constantly on the increase. A broad analysis of studies concerning GT is included in the work [13]. These authors reviewed 657 Scopus articles published in the last 10 years in this field to analyze the directions and areas of research related to GT. They conclude that in the past the purpose of analyzing GT data mainly focused on describing and diagnosing research and business trends, while now the purpose is shifting toward forecasting. Due to the researchers recognizing this potential, areas for the application of predictions from GT data include medical science, the economy and business and now even encompass sociology, politics and law. They noted that in the performance of forecasting, additional analyses, such as sentiment and cross-correlation analyses, will be required.

In a further part of this literature review, we focus on what was new following that period. Therefore, for this study, the latest articles from the Scopus database, from 01.01.2017 to 19.09.2020, were analyzed. We yielded 1,877 papers as a result of the query ''google and trends'' in the title, abstract or keywords. Incomplete studies (e.g. an abstract was missing) or those which had weak relevance to the Google Trends service were removed from the set of articles. Finally, the analysis dataset consisted of 1,012 papers.

A Text Mining approach was used to analyze such a large volume of literature. We analyzed the keywords from articles (using TM algorithms) to discover how they coexist in the literature related to Google Trends. Article keywords are determined by the author and are intended to well reflect the subject and characteristics of the research. Cluster extraction made it possible to identify all similar paragraphs in a large collection of the author's keywords. The algorithm compares text segments (author's keywords from each article) and groups the most similar ones into clusters. The algorithm clusters data into different groups based on a similarity measure. The final output (Table 1) is a list of clusters along with their sample members (keywords).

The most numerous group are studies dedicated to coronavirus, which seems to be an obvious trend at the time of this

particular pandemic. Research in the field of medical science also covers cancer topics and the epidemiology of many diseases. Moreover, the latest scientific trends also include a large amount of research in the field of the prediction of many economic phenomena, such as the forecasting of tourism demand or tourist arrivals, bitcoin price movement and stock prediction, etc.

Analyzing in more detail the literature on the use of GT data for predictive purposes, attention should be paid to the variety of domains in which researchers manage to find applications for them. Table 2 presents the arbitrarily selected studies from 2017 to 2020 to illustrate the multi-domain potential of SVI data for predictive purposes. In the entire analyzed research papers dataset, only one study was dedicated to the real estate market topic. Venkataraman *et al.* [1] examined whether internet search intensity, captured by Google's SVI, predicts house price changes in an emerging market like India. The study covers four large metropolitan cities over 24 quarters between 2011 and 2016. Developing countries such as India are characterized by the high cost and low availability of technology. This limits its usage only to a privileged few. Despite this, their results showed that SVI strongly predicts future house price changes in the Indian real estate market.

Although Table 1 and Table 2 present a very wide application of SVI data for prediction in many areas, the lack of research dedicated to the real estate market is clearly visible.

C. RESEARCH OF GOOGLE TRENDS DATA FOR FORECASTING IN HOUSING MARKET

Through snowballing technique, we also found several publications on this topic from an earlier period, which were neither covered in the study [13], nor our literature analysis of papers from the Scopus database, and take into account English and US housing market specificity (see Table 3). The general findings of these studies are that sellers and buyers contribute significantly to the internet search volume.

In the study by [39], a forecast model for house prices was created. Initially, the model considered a wide range of search terms (including ''house prices'', ''buy house'', ''sell house'', ''mortgage'' and ''estate agents''). Ultimately, the search term ''estate agents'' was chosen as it was much more stable and correlated with both house prices and housing transactions. The dependent variable in the model was the monthly house price growth. The house price growth balances from the Home Builders Federation (HBF) and the Royal Institution of Chartered Surveyors (RICS) are both used in the model as independent variables.

In turn, the study [35] showed that GT data for selected search terms, i.e. ''house for sale'' and ''rentals'', significantly influenced the improvement of the prediction precision of alternative models. The improvement of the precision of models with GT data oscillated in the range of 2.5% - 8%, depending on the selection of queries and the initial date of prediction. However, this study showed the limited utility of the 'mortgage' query in HPI modelling.

In the study [34] the authors started with estimating the baseline model to predict the present home sales using only the past home sales and the past HPI. The past home price and sales were highly correlated with the current home sales. Next, they examined various search indices related to the real estate market and found two categories, ''real estate agencies'' and ''real estate listings'', to best predict the present sales. They revealed that the best predictors of the present home sales are the current index for ''real estate agencies'' as well as its one-quarter and two-quarter lags.

In the paper [2] authors focused on the cross-sectional differences in search intensities among different metropolitan statistical areas (MSAs) and not on search intensity in time series. They explored whether online search intensity for queries that include the words 'real estate' or 'rent' help predict home prices. They investigated the extent to which future cross-sectional differences in home price changes are predicted by online search intensity in prior periods. Their findings suggest that on average, higher or lower abnormal search intensity for real estate in a particular city precedes positive and negative abnormal home price changes for that city relative to the overall U.S. housing market, respectively.

The predictive model created by [12] implied that (i) house sales at (t-1) are positively related to house sales at t, (ii) the search index on ''rental listings & referrals'' was negatively related to sales, (iii) the search index for ''real estate agencies'' was positively related to sales, (iv) the average housing price was negatively associated with sales.

The findings of [42] confirmed the suggestion that searches for the keyword ''foreclosures'' are strongly correlated with actual forecastings in the United States, and at the same time, assumed that search trends can be used as an early warning system for problems in the US real estate market.

Thus, from an exploration of the extant literature devoted to the study of the possibilities of Google Trends data for the analysis and forecasting of various economic and social phenomena, we can conclude that firstly, most of the existing research in the field of the real estate market is based on hedonic-based regression approach to the study of trends and relationships between variables. Secondly, in recent years, more and more works use machine learning algorithms as an effective and promising tool for building regression and performing classification to predict the selected parameters of the studied phenomenon (especially the capital market research domain, which is the most advanced in this). At the same time, there are practically no studies in which hedonic-based regression and machine learning classification algorithms are used comprehensively for mutual enrichment. Thirdly, most housing market-related research (i) include Google Trends keywords as additional parameters to the models which are based on existing determinants of housing price; and (ii) analyze the predictive power of Google Trends data only as an enhancer of the traditional model's accuracy (Table 3). Fourthly, extant literature testifies to the successful results of using Google Trends keywords as the basic and only

TABLE 2. The overview of the most recent studies from variety of disciplines with usage google trends data for forecasting.

Domain	Author	Topic and research results	Main research methods
MEDICAL SCIENCE	[14]	Tracking queries in the Google search engine gives the possibility of **forecasting tuberculosis** case numbers.	seasonal autoregressive moving average model and neural network models using 5-fold cross-validation
	[15]	Identifying variables for **forecasting seasonal influenza** (ILI) **in South Korea** and building prediction models for ILI activity based on GT data.	autoregressive integrated moving average, including exogenous variables (SARIMAX) models, cross-correlation analysis
	[16]	They created the GT data-driven model **to forecast the number of dengue fever cases**.	ARIMA and ARIMAX models
	[31]	The study demonstrates the feasibility of accurate, real-time **influenza-like illness (ILI) incidence predictions** in the Netherlands using Google search query data.	Lasso regression, cross-validation
BUSINESS AND ECONOMY	[5]	Explaining and predicting the regional **unemployment rate in Romania** at the county level using GT data.	Granger causality, panel data model
	[19]	The study investigates the possibility of **predicting total tourist arrivals to four Austrian cities** (Graz, Innsbruck, Salzburg and Vienna).	autoregressive moving average model classes, mixed data sampling (MIDAS) model class, autoregressive distributed lag (ADL) model class
	[1]	Exploring whether internet search intensity, as captured by Google's search volume index (SVI), **predicts house price changes** in an emerging market like India.	the estimation of variables, regression
	[20]	Forecasting **consumer behaviour for the fashion industry** from GT data**.**	singular spectrum analysis, ARIMA model, hybrid neural network model
	[21]	Using data mined from Google Trends, the **prediction of the unemployment rate** among Canadians between 25 and 44 years of age was performed.	the estimation of variables, regression
	[41]	Google Trends data are a **strong predictor for future employment growth** in the United States over the period 2004–2019 at both short and long horizons.	variable selection mechanisms (soft thresholding variable selection), non-linear models such as Random Forests
	[30]	Using Google Trends search query volume data for **tourism demand forecasting** at high spatial detail.	the estimation of variables, regression
CAPITAL MARKETS	[22]	The authors showed that fear of the coronavirus – manifested as excess SVI of GT – represents a timely and valuable data source for **forecasting stock price** variation around the world.	Pearson's correlation, heterogeneous autoregressive (HAR) model
	[25]	The study uses Twitter and Google Trends to **forecast the short-term prices of the primary cryptocurrencies.** The study showed that a combination of GT data and general negative sentiments (including weighted sentiments) was the most powerful predictors.	support vector regression, stochastic gradient descent, gradient boosting model, multilayer perceptron neural network, least squares linear regression, AdaBoost, Bayesian ridge regression, decision tree, ElasticNet, and Hybrid, which was the mean of all of the models
	[26]	They created **models to forecast weekly and monthly volatilities of S&P 500 index.**	hybrid models based on artificial neural networks with multi-hidden layers
	[27]	GT data were taken into consideration for improving stock prediction. Research results proved that GT can help in **predicting the direction of the stock market index**.	back propagation neural networks (BPNN), improved sine cosine algorithm (ISCA)
	[28]	The proposal of the **model on the bitcoin transaction network for forecasting the bitcoin price movement.**	partial differential equation (PDE) model on the bitcoin transaction network
SOCIO-PSYCHOLOGICAL	[17]	They utilized Google Trends search volumes for behavioural **forecasting of national suicide rates in Ireland** between 2004 and 2015.	vector autoregression and neural network autoregression
	[18]	They **built the model which was able to predict the number of suicides** (weekly, for six subgroups, defined by gender and age) in Hong Kong.	Poisson regression models
	[29]	Based on GT data, they created the model for **predicting state-level alcohol-induced death (AICD), drug-induced death (DICD), and suicide rates**.	cross-validation, estimation of coefficients, regression
RAW MATERIALS MARKETS	[23]	They constructed **a predictive G-trends-based model providing for profits from precious metals**, such as Gold, Palladium, Platinum and Silver.	econometric model, autoregressive distributed lag (ARDL) model
	[7]	They created the internet search-driven **model** (GT data) **to forecast crude oil prices**.	estimation, ARMAX model
	[24]	Based on daily search query data from Google Trends, it is possible to **predict oil price movements** and an increase in volatility in the trading days that follow.	covariance tests, structural vector autoregressive (SVAR) model, causality, and volatility spillover models

determinant in forecasting models [31], [32], [25], in several research domains. However, so far no research has been conducted on the ability of Google Trends data to *solely* predict housing price changes (shift direction) based on a

TABLE 3. Application of google trends data for forecasting in housing market-related research.

Author	Topic and research results	Main research methods	Determinants
[39]	The model with the internet search term variable was able to predict a significant proportion of growth in the United Kingdom's house prices.	A regression model	*Google Trends keywords*: "estate agents" *Traditional*: past house price value, HBF, RICS
[35]	The findings confirm the significant impact of GT data on the improvement of precision in the estimation of HPI changes for 2004–2014 in Great Britain.	Estimation of autoregressive models	*Google Trends keywords*: "house for sale"', "rentals", "mortgage" *Traditional*: past sales volume, past HPI value
[34]	They discovered that HPI is strongly predictive of future housing market sales and prices in the US housing market.	A seasonal autoregressive (AR) model	*Google Trends keywords*: "real estate agencies", "real estate listing", "quarterly sales" *Traditional*: past sales volume, past HPI value
[2]	Their findings showed that abnormal search intensity for real estate in a particular city in the US can help predict the city's future abnormal housing price change.	A regression specification model, Granger causality tests and panel vector autoregressions	*Google Trends keywords*: "real estate" and "rent" for U.S. metropolitan statistical areas, *Traditional*: HPI
[12]	They discovered that the search index for "real estate agencies" was the best predictor for contemporaneous house sales in the US housing market.	Econometrics models that include the Google Trends variables.	*Google Trends keywords*: the search index on "rental listings & referrals" and "real estate agencies" *Traditional*: house sales
[40]	The empirical results show that all models augmented with Google data, combining both macro and search data, significantly outperform baseline models which abandon internet search data. Models based on Google data alone outperform the baseline models in all cases.	Three groups of models: baseline models including fundamental macro data only, those including Google data only and models combining both sets of data. One-month-ahead forecasts based on VAR models. Granger-causality tests.	*Google Trends keywords*: general market indices (e.g. "commercial and investment real estate"), and sector-specific indices search (e.g. "jones lang lasalle" or "loopnet") *Traditional*: prices, transactions

machine learning *classification* approach and with the context of the *Polish* market.

To fill these gaps, in this study we *aim* to extend the understanding of the Google search engine query volume extra predictive power of future housing price change, in addition to all existing housing price determinants, by (i) the introduction of a complex research approach that combines the advantages of the complementary use of cross-correlation analysis and machine learning classification algorithms; (ii) applying multi-class HPI values classifier which allows predicting the housing price increase, decrease or relative stability; (iii) exploiting the Google user search volume that relates to interests in both 'real estate' and 'credit to buy real estate'; (iv) the evaluation of introduced approach in the context of the Polish real estate market.

III. METHODOLOGY

To tackle the research challenges described above, we propose a comprehensive approach, which includes the following steps (Figure 2): (1) Data collection and preparation; (2) A cross-correlation analysis of the presence of dependence between quarterly HPI values and the Google user search volume (SVI), which are related to each individual keyword from two selected perspectives – (i) real estate and (ii) credit to buy real estate; (3) Exploring the machine learning classification algorithms for predicting real estate price change based on Google user search volume (SVI) in the context of the Polish market.

A. DATA COLLECTION AND PREPARATION

Following the aim of this study, two sources were used for *collecting* the dataset. The *first* data source is the https://tradingeconomics.com/ webpage,[1] which allowed us to collect the data about the House Price Index in Poland in the last ten years, but without affecting the pandemic period (i.e., between 01.01.2011 and 31.12.2019). HPI is freely downloadable and is presented in a quarterly timestamp, and captures price changes of all kinds of residential property purchased by households.[2]

The *second* data source is the Google Trends tool (https://trends.google.com/trends). Our sample selection process in GT was based on four parameters: (i) search engines, (ii) country, (iii) time range and (iv) search terms (keywords).

[1] https://tradingeconomics.com/poland/housing-index

[2] In Poland, the House Price Index measures residential property market prices. The HPI captures price changes of all kinds of residential property purchased by households (flats, detached houses, terraced houses, etc.), both new and existing. Only market prices are considered, self-build dwellings are therefore excluded. The land component of the residential property is included (https://tradingeconomics.com/poland/housing-index)

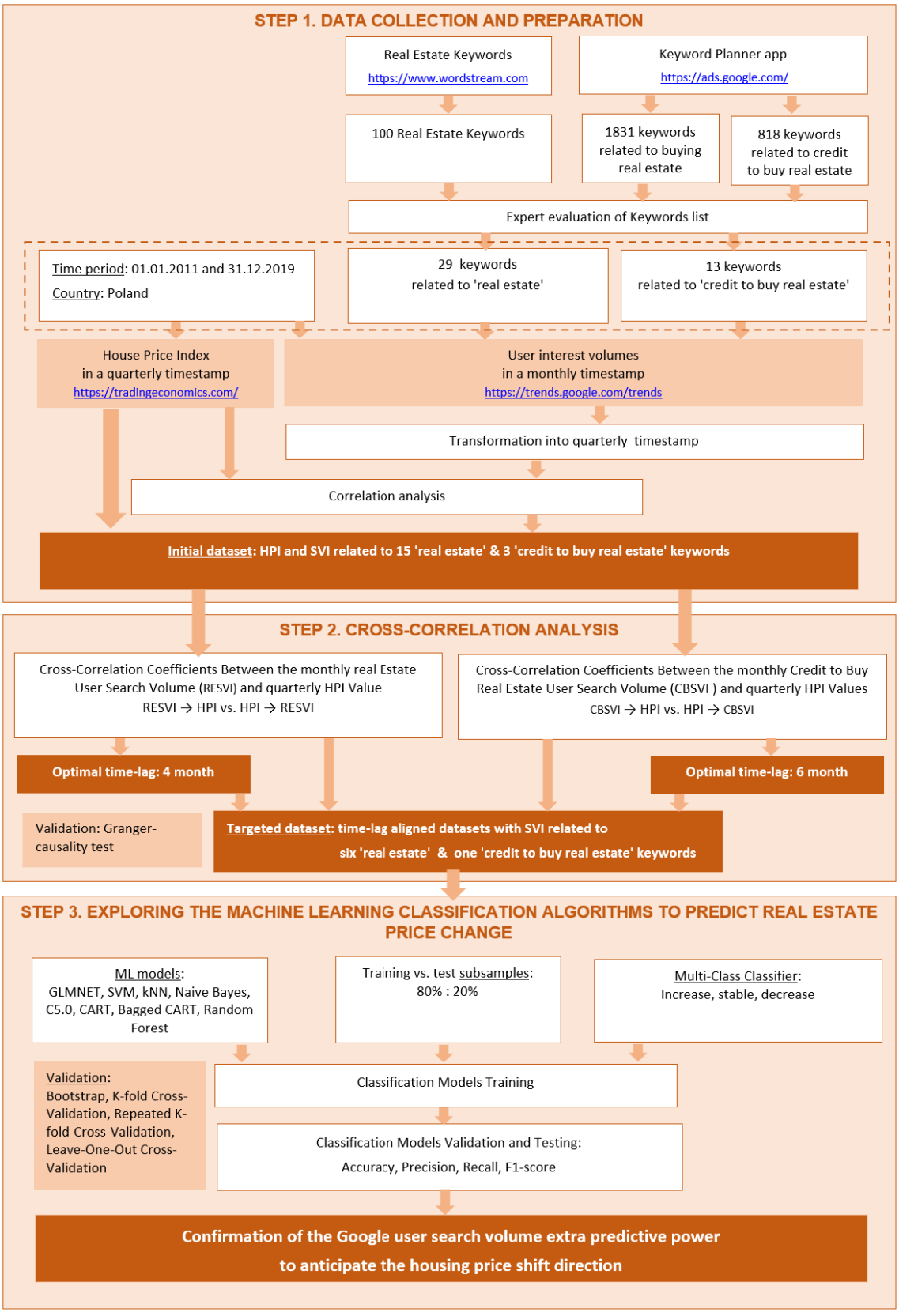

FIGURE 2. Step-by-step model of research methodology.

Queries were collected from all five proposed specialized *search engines*: Web, Image, News, Google Shopping, and YouTube Search. Poland was indicated as a *country* selection parameter. Synchronized with House Price Index data, the period between 01.01.2011 and 31.12.2019 was defined as a *time rang*e. User interest volumes (SVI) in Google Trends are presented with a monthly timestamp.[3] The retrieval date for both data sources was 15.09.2020.

In the further part of the research, a difference in timestamp between SVI and HPI data was proposed to be used in two ways: (1) the monthly timestamp of Google Trends data can be aligned to HPI data by *converting* monthly SVI values into a *quarterly* timestamp. Then, these quarterly SVI values can be used at the initial stages of the study (i) to confirm the assumption about the feasibility of this research (checking the presence of significant correlation between HPI and SVI time series); and (ii) to select sufficiently relevant keywords, which determine the SVI data used at the cross-correlation analysis stage; (2) the *monthly* timestamp of Google Trends data can be used in its *original* form (without quarterly alignment) with the aim (i) to improve the accuracy of cross-correlation analysis and the optimal value of the time-lag search; and (ii) to select highly relevant keywords, which determine the SVI data used at the stage of testing the predictive potential of the Google user search volume in the HPI context.

In order to form a set of *search terms (keywords)* to collect the Google Trends dataset, we were guided by the goal of identifying the Google users interest in (1) offers to buy various types of *real estate*, contextually connected with the HPI concept (i.e, houses, homes, flats and apartments); and (2) the possibility of using *credit* to make such a purchase. To collect the initial search keywords list the following resources selected for this study, namely: (i) Keywords from the most popular suggestions for real estate domain (generated from utilized latest Google search data)[4]; (ii) the Keyword Planner app at https://ads.google.com/, which generates ''keyword ideas'' close to the keywords specified in the app's search bar. As a result, the initial keywords sample was obtained, which includes: (1) 100 keywords from Real Estate Keywords (Appendix A); (2) 1831 keywords concerning buying various types of *real estate* (Appendix B); (3) 818 keywords related to the possibility of using *credit* to buy *real estate* (Appendix C).

To *select* the most relevant *search keywords* from the initial set, the following three approaches were applied: (i) an expert approach; (ii) a Pearson correlation coefficient; and (iii) a cross-correlation analysis. On the data collection and preparation step, the first two approaches were realized. For this, two independent experts (one an academic researcher and the other a real estate practitioner) discussed and developed the following *criteria list* to select keywords that are relevant to the research objectives: (i) keywords should refer only to buying (not selling or renting) real estate; (ii) keywords should not be repeated; (iii) keywords should not contain names of cities or regions of Poland; (iv) keywords should not give zero search results; (v) keywords can be both in Polish and English, since we admit the fact that foreign investors can also enter the Polish real estate market. As a result of the first round of the relevant *search keyword selection*, 42 keywords (29 concerning buying various types of *real estate* and 13 keywords regarding the possibility of using *credit* to buy *real estate*) were selected (Appendix D). Based on the selected search keywords list, the initial dataset of *user interest volumes* (i.e. SVI) from the Google Trends tool was collected. When retrieving data, keywords with more than one word should be written in the GT search term box without quotes.

In the *preparation* phase of the collected dataset, the following three stages were realized. *Firstly*, to decrease the sample variance, the GT and HPI dataset values were *normalized* using the linear technique.[5] The highest interest in the search query is expressed by 1, whereas a lack of interest or insufficient data is expressed by 0. *Secondly*, Google Trends user interest volumes were converted from a *monthly* to quarterly timestamp. As a converting approach, the summarization function was selected. *Thirdly*, the correlation analysis was performed to identify the relationship between quarterly values of (i) HPI and (ii) the Google *user search volume* (SVI) related to each individual keyword from two selected perspectives (*real estate and credit* to buy *real estate*). In the next step of the study, only SVI data related to the terms for which the Pearson correlation coefficient was greater than 0.5 and statistically significant were analyzed (second approach of the relevant *search keyword selection*).

B. CROSS-CORRELATION ANALYSIS

To answer the first research question (RQ1), the next step of our study was the process of confirming the assumption of the *presence of a correlation between the changes in the user search volume and prices in the real estate market with a particular time-lag.* Because the Pearson correlation coefficient does not provide information about directionality between the two time-series, to identify the time shifts between real estate prices and Google user interest movements, the time-lagged Pearson cross-correlation coefficient could be applied[6] [4]. The time-lagged Pearson cross-correlation coefficient method consists of computing the ordinary Pearson correlation with the two-time series shifted by a variable time interval. The lag σ refers to how far the series is offset, and its sign determines which series is shifted [89]. In a cross-correlation in which the direction of influence between two time-series is hypothesized or known,

[3] The SVI numbers represent the search interest relative to the highest point on the chart for the selected region and time. A value of 100 is the peak popularity of the term, whilst a value of 50 means that the term is half as popular (https://trends.google.com/trends)

[4] https://www.wordstream.com/popular-keywords/real-estate-keywords

[5] $x' = \frac{(x - x_{\min})}{(x_{\max} - x_{\min})}$

[6] $r(\sigma) = \frac{\sum_{t=1}^{n} (Q_t - \bar{Q})(T_{t+\sigma} - \bar{T})}{\sqrt{\sum_{t=1}^{n} (Q_t - \bar{Q})^2} \sqrt{\sum_{t=1}^{n} (T_{t+\sigma} - \bar{T})^2}}$, where, are the $\bar{Q}$ and $\bar{T}$ is the sample averages of the two time series (in our case HPI and SVI respectively), σ is a time-lag.

the influential time-series is called the ''input'' time-series and the affected time-series is called the ''output'' time-series. The application of cross-correlations in the context presented herein infers that the *input* time-series refers to Google user interest (SVI) and the *output* time-series refers to real estate market prices (HPI).

Four potential theoretical cases can be defined to explain the cross-correlation between the volume of Google user interest and real estate market prices: (C1) Google SVI and real estate market prices are positively correlated with a positive time-lag. An increase in Google SVI causes an increase in real estate market prices; (C2) Google SVI is positively related to real estate market prices with a negative time-lag. An increase in real estate market prices causes an increase in Google SVI; (C3) Real estate market prices respond negatively to Google SVI with a positive time-lag. An increase in Google SVI causes a decrease in real estate market prices; (C4) A decline in real estate market prices is followed by an increase in Google SVI (*negative time-lag*).

To confirm the presence of both (i) a *correlation dependence* between the changes in the user search volume and real estate market prices, and (ii) a particular *time-lag* for such a correlation, the following four stages were realized:

(1) Because data on the user search volume (SVI) are presented in the dataset with a monthly timestamp, and the HPI index - with a quarterly timestamp, it was decided to study the existence of a relationship between (i) SVI in each month of the particular quarter and (ii) the quarterly HPI values. Hence, firstly, for each of the group of search keywords (i.e., real estate and credit to buy real estate) three subsamples were formed of average SVI - corresponding to the 1st, 2nd and 3rd months of the quarter.

(2) Indicators of cross-correlation coefficients with a time-lag range from 1 to 9 months were calculated to explore all four potential theoretical cases (C1-C4) for each created subsample.

(3) The optimal value of the time-lag, which allows for asserting the existence of a significant correlation between the change in SVI and the subsequent change in the HPI values, was chosen.

(4) Cross-correlation coefficients between each individual keyword and the HPI index with a chosen optimal time-lag were calculated. As a result, (i) individual optimal time-lag values were identified and (ii) a list of search keyword indicators that have the greatest significance in predicting HPI changes, was selected (third approach of the relevant *search keyword selection*). This list of search keywords determines a targeted dataset structure.

C. EXPLORING THE MACHINE LEARNING CLASSIFICATION ALGORITHMS TO PREDICT REAL ESTATE PRICE CHANGE

To answer the second research question (RQ2), the process of exploring the *predictive potential of the Google user search volume (expressed by SVI) using the machine learning classification approach in the context of HPI in the Polish market* is accepted as the next step of our research. To realize this step, the following *three* stages of analysis were performed:

Stage 1. Eight popular machine learning (ML) classification algorithms were selected in our study: (i) *Linear methods:* Regularized Regression (GLMNET); (ii) *Non-Linear methods*: SVM, kNN, Naive Bayes; (iii) *Trees and Rules*: CART; and (iv) *Ensembles of Trees*: *C5*.0, Bagged CART, Random Forest.

It is well known that multiple linear regression is often used to estimate a model for predicting future responses. Due to that ordinary least squares (OLS) regression not often performing well with respect to both prediction accuracy and model size, several *regularized regression* methods were developed in the last few decades to overcome these flaws of OLS regression, starting with ridge regression [62], followed by lasso method [61], and more recently the elastic net [60], [58], [59], [63]. Ridge regression and the lasso are regularized versions of the least squares regression using penalties on the coefficient vector. Elastic net was developed to reach a compromise between the lasso and ridge regression for improving the prediction accuracy. Elastic Net was included in the variable selection process in a fitted Support Vector Regression (SVR), forecasting the yearly U.S. Real Housing Price Index [74]. In our research, the Lasso and the elastic-net regularized generalized linear model (*GLMNET*), introduced by [53], was used [54]. In GLMNET, each parameter is optimized by the minimization of the objective function; whereas, the remaining parameters are fixed. GLMNET implements optimization for each parameter of the model and the optimization process is continuously performed. As the key parameters to tune the accuracy of the GLMNET model, the parameters alpha (Mixing Percentage) and lambda (Regularization Parameter) are used [52], [55]–[57], [73]. In our research, the optimal model was built with *alpha* =0.55 and *lambda* = 0.1570234.

*Support vector machin*e (SVM) is a non-linear method, which was first introduced by [64] for classification problems and is intended to find the optimal separating hyperplane among the classes by maximizing the margin between them. Usually, when using SVM, two problems are confronted: how to choose the optimal input feature subset for SVM and how to set the best kernel parameters. The SVM method has been proven to be very effective for addressing housing market price classification and prediction problems [43]–[45]. Although, like neural networks models, they are often marked by users as ''black boxes'', lacking clarity of the resulting solutions [46]. In our research, we will implement a Support Vector Machine algorithm with Radial Basis Function Kernel [47]. Tuning parameter *'sigma'* was held constant at a value of 0.00453203, regularization parameter *C* changed to be 0.25. *Kernel cache* and *maximum iteration* values are 200 and 10000000.

The *K-Nearest Neighbors* (KNN) is a non-parametric non-linear method used for classification and regression. By defining a special number k in the total data set, the average/modes classes of the nearest neighbors are obtained and

the new object is assigned to the nearest class to its neighbors. The distances of the new object with its neighbors can be calculated with functions such as Euclidean, Manhattan, Minkowski, and Chebyshev [90]. The KNN algorithm is one of the methods used for classification analysis, but the last few decades it has also been used for prediction, including real estate market price [51], [48]–[50]. For this method, the choice of the k value is extremely important. In our research, the best prediction result observed is that in which k is equal to 9.

Naïve Bayesian is a statistical non-linear learning algorithm based on Bayes' rule to compute joint probability with the assumption about conditional independence amongst the attributes. In Naïve Bayesian first the dataset is divided into independent classes and then the probability distribution for each attribute of each class is calculated. For classification, the Naïve Bayesian finds the probability for the unknown in any given class and selects the class with the highest probability. The strength of Naïve Bayesian classifier, as a powerful probabilistic model, has been proven for solving house sale (resale) price classification and prediction tasks effectively [65], [45], [66]. Tuning parameters in this research are: using a kernel density estimate for continuous variables versus a Gaussian density estimate; '*Laplace smoother'* was held constant at a value of 0; '*adjust'* parameter to bandwidth of the kernel density was held constant at a value of 1.

One noticeable advantage of decision tree-based models is that they are scalable to large problems and can handle smaller datasets than many other ML models. *Classification and Regression Tree* (CART) models referred to trees and rules algorithms, which are a set of if-then (split) rules and are permitted prediction or classification of cases [66]. A CART model is referred to as the regression-type model and was developed by [70] to predict the value of continuous variables from a set of continuous and/or categorical predictor variables. CART models use a binary tree to recursively partition the predictor space into subsets in which the distribution of predictor variable y is successively more homogenous. CART advantage is its ability to select the most discriminatory features and that the classification is done with fewer calculations [77]. From the previous studies' results it is known that the CART model provides better housing price prediction compared to regression models, also effectively supporting real estate market participants decision-making, also based on crime occurrence [65], [67]–[69]. In our analysis, we needed to control the modelling process by the complexity parameter (*cp*), which imposes a penalty to the tree for having two many splits. The higher the *cp*, the smaller the tree.

Ensemble modelling is a process that is used for prediction either (i) for several different modelling algorithms or (ii) for different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data. The motivation for using ensemble models is to reduce the generalization error of the prediction [76]. Ensemble decision trees methods combine several decision trees to produce better predictive performance by the principle that a group of weak learners come together to form a strong learner. The main techniques to perform ensemble decision trees are *Bagging* and *Boosting*. *C5.0* is a new decision tree algorithm developed based on C4.5 by Quinlan [79]. It includes all functionalities of C4.5 and applies a bunch of new technologies, among them (1) the branch-merging option for nominal splits is the default; (2) misclassification costs can be specified; (3) *boosting* and cross-validation are available; and (4) the algorithm for creating rule sets from trees is much improved [78], [71], [65], [72]. Because the C5.0 algorithm is quite new, only a few studies confirm the possibility and effectiveness of its use to predict housing price or housing defaults [80], [72]. There are three meta parameters for tuning the C5.0 that were used: $trials = 1$, $model = rules$ and $winnow = FALSE$.

Bagged CART is an ensemble *bootstrapped* aggregation method that fits many trees to bootstrap-resampled versions of the training data to build independent prediction models, and then combines them using an averaging technique. Because this technique takes many uncorrelated trees to make a final model, it reduces error by reducing variance [84], [81]–[83]. Designing Bagged CART ensemble learners has been recognized as one of the significant trends in the field of big data analysis and last years begin to demonstrate a high level of competitiveness in predicting house prices [49], [85].

The *Random Forest* algorithm is one of the most popular machine learning models, based on the *bagging* and random subspace methods; it can perform both regression and classification tasks, just like a Decision Tree algorithm. These algorithms are very often used to address the problem of forecasting housing prices, price analysis and real estate decisions [86], [67], [48]. The idea of bagging is to construct an ensemble of learners, each trained on a bootstrap sample obtained from the original dataset. The prediction of the ensemble is constructed from the separate decisions by majority voting (classification) or averaging (regression). It has been shown that bagging can reduce the variance in the final model when compared to the base models and can also avoid overfitting [87], [83]. For Random Forest model tuning, we needed two parameters: (i) mtry, which controls the number of predictor variables randomly sampled to determine each split; (2) ntrees, which controls the total number of independent trees [91]. In our models, we considered that *mtry* tuning length is equal to 15 (from 123) with the optimal value used for the model 88. Predictions were made on the 20% tuning data set from subsets of increasing numbers of *trees* (1000 trees, 2000 trees, ..., 2500 trees).

Supervised learning was used for all eight selected ML algorithms. This refers to the fact that the samples were divided into the (i) *training* part, allowing the models to "learn" expected output values (Y) based on the input-provided values (X), and the (ii) *test* part, where the ability of the model to indicate was verified with previously unknown Y values based on X values not coming from the

Z training set [32]. The main training model *setting* is: (i) as an output variable the HPI values were selected; (ii) as an input variable the SVI values were assigned; (iii) training and test subsamples were selected from the targeted dataset with the proportion 80% **:**20%; (iv) to perform the output variable classification, three classes were developed (Table 4). Such classification rules were adopted based on the definition of price stability established by the European Central Bank[7] in the following edition: house price *stability* could be defined as a change in the House Price Index with the safety margin of $+/-$ 2% around HPI value which equals 100. The HPI values above and below the accepted safety margin will refer to the classes of *increase* and *decrease* of House Price Index accordingly.

TABLE 4. Multi-class classification rules.

Class labels	HPI values
Increase	>102
Stable	>=98 and <=102
Decrease	<98

Stage 2 . *Validation* of the trained predictive models. Model validation helps ensure that the model performs well on new data, and helps select the best model, the parameters, and the accuracy metrics. As a result, well-performing models were selected.

Stage 3 . *Testing* the selected machine learning models to estimate the predictive potential of selected machine learning models to anticipate the direction of real estate market price changes under the influence of changes in the Google user search volume (SVI).

This step of the analysis was conducted in the R software environment (the developed code is available on the Github repository). HYPERLINK "https://github.com/TextAnalitics/Prices-Prediction-for-Real-Estate-Market/blob/main/Mashine%20Learning%20Prediction.R" repository

D. RESULTS VALIDATION

The stages of validation of the obtained results consist of: (1) an assessment of the degree of consistency of the findings of this research with the conclusions obtained in the previous literature; (2) using the implementation of several model validation techniques: Granger-causality test (RQ1); Bootstrap, K-fold Cross-Validation, Repeated K-fold Cross-Validation, Leave-One-Out Cross-Validation (RQ2).

IV. EXPERIMENTAL RESULTS

A. DATA COLLECTION AND PREPARATION

In accordance with the developed research methodology, the initial dataset was collected from two sources and, as a result, consists of two main parts. The *first* part containing 36 House Price Indices values, presented price changes of all kinds of residential property purchased by households in Poland in the period between 01.01.2011 and 31.12.2019, presented in a quarterly timestamp (from Q1 2011 to 4Q 2019).

The *second* part of the initial dataset contains *user interest volumes,* including both "real estate" and "credit to buy real estate" user interest volumes, and collected from Google Trends tool using 42 selected *by experts* keywords (totally 42 columns and 108 rows). All two parts of the collected dataset are presented in Appendix E.

In order to increase the relevance of data, used in the subsequent stages of our research as independent variables, the collected dataset was preprocessed and additionally reduced. As a criterion for the relevant data selection, Pearson correlation coefficient between quarterly values of (i) HPI and (ii) the Google *user search volume* (SVI) related to each individual keyword from two selected perspectives (*real estate and credit* to buy *real estate*) was used. Following our methodology, for calculating the Pearson correlation coefficient values, (i) Google Trends and House Price Index dataset values were *normalized*; (ii) Google Trends user interest volumes were converted from a *monthly* to quarterly timestamp. As a result of the data *preparation* step, the initial dataset was reduced to 18 *search keywords*: 15 concerning buying various types of *real estate* and 3 keywords related to the possibility of using *credit* to buy *real estate* (second approach of the relevant *search keyword selection*). The final list of *search keywords,* which determines the initial dataset structure, with information about the Pearson correlation coefficient between quarterly HPI values and Google Trends data, related to SVI using individual keywords from two selected perspectives, is presented in Table 5.

The descriptive statistical of *the initial dataset* (with quarterly timestamp) is presented in (i) Appendix F (HPI values) and (ii) Appendix G and Appendix H (SVI values related to selected 18 search keywords from the perspectives of *real estate* and *credit* to buy *real estate*).

B. CROSS-CORRELATION ANALYSIS

We summarize the results of our experiments based on our research questions. The first research question (RQ1) is aimed at extending the understanding of the *nature of dependencies between the Google user search volume (SVI) and Polish real estate market prices (HPI)* and *their dependent changes*. Following the proposed methodology, we performed cross-correlation analysis for the average SVI values related to (i) *real estate* and (ii) *credit* to buy *real estate* separately.

During the cross-correlation analysis of the *'real estate'* related user search volume (RESVI), the following findings were obtained: (1) the real estate Google user search volume and real estate market prices (HPI) changes are *strongly correlated* (the correlation coefficient between these two indicators in the quarter timestamp is equal to 0.77); (2) the

[7] https://www.ecb.europa.eu/mopo/strategy/pricestab/html/index.en.html

TABLE 5. The list of search keywords in the initial dataset.

Keywords	Correlation Coefficient	Keywords	Correlation Coefficient
Real Estate Users Interest			
"houses for sale"	0.6808	"mieszkanie na sprzedaż"[8]	0.8211
"new home"	0.5022	"domy na sprzedaż"[9]	0.7549
"property for sale"	0.6102	"mieszkania na sprzedaz"[10]	0.6310
"morizon domy"[11]	0.7643	"nieruchomości na sprzedaż"[12]	0.6133
"commercial real estate"	0.5493	"mieszkania na sprzedaz olx"[13]	0.7837
"otodom domy na sprzedaz"[14]	0.6837	"mieszkania na sprzedaż olx"[15]	0.8677
"domy na sprzedaz olx"[16]	0.7337	"willa na sprzedaż"[17]	0.6087
"dom na sprzedaz"[18]	0.6995	-	-
Credit to Buy Real Estate Users Interest			
"kredyt na mieszkanie"[19]	0.5347	"kredyt na dom"[20]	0.6602
"wkład własny mieszkanie"[21]	0.5667	-	-

approach which consists of analysing the presence of the time-lag between the changes in *monthly* RESVI and *quarterly* HPI values *confirms its appropriateness* since it allows more accurate tracking of the reaction of these indicators to mutual changes; (3) the RESVI and HPI indicators are *positively correlated* with a *positive* time-lag. This is evidenced by the fact that according to the results of our experiments, (i) the correlation coefficients are positive and (ii) their values in cases with positive time-lags are always greater than the corresponding negative ones (Table 6). This allows the suggestion that today's user search volume anticipates and effects the change in real estate market prices typically in *four* (correlation coefficient equal to 0.84) or, in the more distant future - *seven* months (correlation coefficient equal to 0.87). That is, the change in user search interest at the beginning of one quarter most significantly effects the change in real estate prices only at the *beginning* of the *next* quarter.

Additional cross-correlation analysis of RESVI values for *individual* search keywords made it possible (i) to identify *individual* optimal time-lags, which vary from 4 to 7 months,

[8] "apartment for sale" (English)
[9] "houses for sale" (English)
[10] "apartment for sale" (English)
[11] "morizon houses" (English)
[12] "real estate for sale" (English)
[13] "apartments for sale olx" (English)
[14] "otodom houses for sale" (English)
[15] "apartments for sale olx" (English)
[16] "houses for sale olx" (English)
[17] "villa for sale" (English)
[18] "house for sale" (English)
[19] "loan for an apartment" (English)
[20] "loan for a house" (English)
[21] "own contribution an apartment" (English)

TABLE 6. Cross-correlation coefficients between the real estate user search volume and HPI values.

Time	RESVI → HPI (positive time-lags)	HPI → RESVI (negative time-lags)
Quarter by Quarter	0.77	0.77
1 month lag	0.79	-
2-month lag	0.74	-
3-month lag	0.76	-
4-month lag	0.84	0.76
5-month lag	0.71	0.71
6-month lag	0.81	0.73
7-month lag	0.87	0.72
8-month lag	0.83	0.65
9-month lag	0.85	0.68

and then (ii) to select the Top-6 of *real estate* search keywords (Fig. 3) that have the greatest effect on the HPI values changing, namely: "*mieszkania na sprzedaż olx*" ("flats for sale in olx"), "*mieszkanie na sprzedaż*" ("flat for sale"), "*mieszkania na sprzedaz olx*" ("apartments for sale olx"), "*domy na sprzedaz olx*" ("houses for sale olx"), "*domy na sprzedaż*" ("houses for sale"), "*dom na sprzedaz*" ("house for sale") (third approach of the relevant *search keyword selection*).

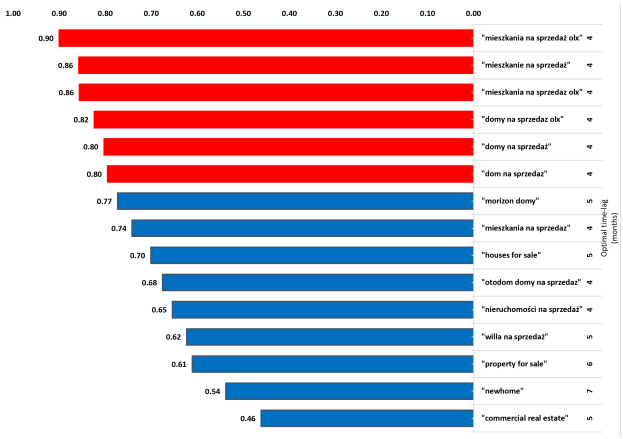

FIGURE 3. Cross-correlation coefficients between RESVI by individual keywords and HPI values.

During the cross-correlation analysis of the average '*credit to buy real estate*' related user search volume (CBSVI), the following findings were obtained: (1) the credit to buy real estate Google user search volume and HPI values are *correlated* (the correlation coefficient between these two indicators in the quarter timestamp is equal to 0.58); (2) the CBSVI values and correspondent market prices are *positively correlated* with a *positive* time-lag. This is evidenced by the fact that according to the results of our experiments, (i) the correlation coefficients are positive and (ii) their values in cases with positive time-lags are always significant and greater than the corresponding negative ones (Table 7). Based on these

TABLE 7. Cross-correlation coefficients between the credit to buy real estate user search volume and HPI values.

Time	CBSVI → HPI (positive time-lags)	HPI → CBSVI (negative time-lags)
Quarter by Quarter	0.58	0.58
1 month lag	0.53	-
2-month lag	0.51	-
3-month lag	0.60	-
4-month lag	0.62	0.49
5-month lag	0.71	0.48
6-month lag	**0.72**	0.54
7-month lag	0.66	0.41
8-month lag	0.61	0.38
9-month lag	**0.74**	0.48

TABLE 8. Cross-correlation coefficients between CBSVI by individual selected keywords and HPI values.

Time-Lag	6-month	4-month	5-month
Keywords	"*wkład własny mieszkanie*"	"*kredyt na dom*"	"*kredyt na mieszkanie*"
Correlation with Time-Lag	**0.71**	**0.63**	**0.54**

results, we can assume that today's CBSVI values are able to anticipate and effect a change in the HPI indicator typically in *six* (correlation coefficient equal to 0.72) or, in the more distant future - *nine* months (correlation coefficient equal to 0.74). That is, the change in user interest at the beginning of one quarter could effect the change in real estate prices only at the *end* of the *next* quarter.

Additional cross-correlation analysis of CBSVI values for *individual* search keywords made it possible (i) to identify *individual* optimal time-lags, which vary from 4 to 6 months (Table 8) and then (ii) to select the one *credit to buy real estate* search keyword that has the greatest effect (0.71) on the HPI values changing, namely: "*wkład własny mieszkanie*" ("own contribution an apartment") with a 6-month time-lag.

To validate the results obtained by the cross-correlation analysis, we applied the Granger-causality test. Generally, the Granger-causality test allows for checking the assumption that the average *user search volume related to* (*i*) *offers on the real estate market and (ii) the possibility of obtaining credit to buy real estate, Granger-causes the value of real estate market prices change*. Keeping the same logic as we had in the cross-correlation analysis stage, we also tried to test Granger causality in two directions. The results of the Granger-causality test are presented in Appendix I. The directions of causality in each test have the following interpretation: for example, RESVI/CBSVI→HPI means that the null hypothesis is that "RESVI/CBSVI does not Granger-cause HPI". The condition for rejecting the null hypothesis is the condition by which the P-value is less than 0.05.

In all our Granger-causality test experiments we observed the same dependencies that we reported in the results of the cross-correlation analysis, namely: (1) Granger-causality in → HPI the direction of the test is much stronger (P-value < Significance level) than the opposite one HPI → (in the majority of cases, P-value > Significance level). Thus, we can conclude that the SVI has a stronger relation to real estate prices, while the real estate price does not have a high degree of causality to the SVI. Similar results were obtained by [38]; (2) The Granger-causality test using a time-lag of 1 to 9 months has demonstrated the presence of (i) a stable causality in the direction up to a 4-month lag, and (ii) weak causality at this lag value in → HPI direction in the case of RESVI (i.e., bidirectional causality). After a 4-month lag testing, the significance of the causality decreases. A similar phenomenon with a lag of 6 months is observed for tests with data relating to CBSVI, but the bidirectional causality with this time-lag has not been diagnosed.

Thus, the cross-correlation analysis step made it possible (1) to confirm the presence of a *correlation* between the changes in the Google user search volume (SVI) and the housing price change (HPI); (2) to establish the *asynchronous* nature of the revealed correlation and identify the optimal size of the time-lag; (3) to select a *targeted dataset* from the initial one, which we will consider relevant for the next study phase; (4) to realize that the Google user search volume for interest in credit to buy real estate is *less correlated* with changes in real estate prices than general information about current offers on the real estate market.

As a result, the *targeted dataset* contains the aligned series of two groups of indicators: (i) House Price Indexes and (ii) normalized values of SVI by seven chosen search keywords (six from *real estate* and one from *credit* to buy *real estate* perspectives, Figure 3 and Table 8). The same 4-month time-lag alignment between the values of two groups of indicators was adopted for the next step of experiments (based on the agreement of the results of cross-correlation analysis and Granger-causality test). This alignment presupposes a shifting of the indicators values by the following *rule*: the SVI value in the first month of each quarter corresponds to the House Price Index value in the next quarter.

C. EXPLORING THE ML CLASSIFICATION ALGORITHMS TO PREDICT REAL ESTATE PRICE CHANGE

The answer to the second research question (RQ2) identifies and improves our knowledge about the *predictive potential of the Google user search volume in the context of HPI shift direction in the Polish market*. To obtain this answer, for our experiments we used a targeted dataset with a total of 280 observations consisting of one output variable (Y) and seven input (X) variables. In the stage of preparing the experiment, all the values of the output variable (HPI) were classified by the rules introduced in Table 4. For each of the eight selected classification models (GLMNET, SVM, kNN, Naive Bayes, C5.0, CART, Bagged CART, Random Forest) the targeted sample was divided into training and test

TABLE 9. Structure of the predictive model's training and test subsamples.

Class labels	Training Subsample	Test Subsample	Total by Class
Increase	11	3	14
Stable	13	2	15
Decrease	5	1	6
Total by Sample	29	6	35

subsamples in a ratio of 80% to 20%. The structure of the subsamples before building the predictive model is presented in Table 9.

Training models were *validated* using Bootstrap, k-fold Cross, Repeated k-fold Cross (k = 10) and Leave-One-Out Cross-Validation algorithms (Appendix J). To *evaluate* the training models, we used two basic indicators: (i) Accuracy as the percentage of correctly classified instances out of all instances, and (ii) Cohen's Kappa coefficient to measure the inter-rater reliability for qualitative (categorical) items. The summary of the results of the evaluation of trained predictive models (*average* from four validation algorithms) is presented in Table 10. These results allow us to select *four* classification models with the (i) highest values of the Accuracy indicator and (ii) satisfactory values of Cohen's Kappa indicators for further testing of the Google user search volume predictive capabilities. Such selected models are GLMNET, C5.0, Bagged CART and Random Forest (highlighted by the red colour in Table 10).

Exploring the performance of the selected predictive models *confirmed* our assumption about the ability of most of these models (except C5.0) to predict changes in HPI values based on the Google user search volume with appropriate (in the range between 0.67 and 1) accuracy. The predictive abilities of machine learning classification models, described by Accuracy, Precision, and F1-score statistics estimation measures, are presented in Table 11.

To evaluate the *generalization* ability of the model, *first,* we detected the selected predictive models overfitting as a difference in prediction performance (accuracy) between the training data and the test data (Appendix L). The results obtained confirmed the presence of the overfitting phenomenon (i.e., the model performs better in the training data than in the test data) for GLMNET and C5.0 models. However, in Bagged CART and Random Forest training models even the maximal accuracy value is not higher than the validation accuracy. This fact proves that using machine learning algorithms for prediction House Price Indices changes based on only Google search activity able to guarantee not accidental results. *Second,* we performed the experiments for the imbalanced training and test subsamples. The received results demonstrated that the accuracy of the predictive models is slightly reduced

TABLE 10. Summary of the evaluation of predictive models.

Algorithms	Min.	1st QR	Median	Mean	3rd QR	Max
Accuracy						
GLMNET	0.33	0.54	0.71	0.74	1.00	1.00
SVM	0.33	0.33	0.50	0.45	0.50	0.50
KNN	0.33	0.38	0.50	0.52	0.50	0.67
CART	0.00	0.33	0.33	0.36	0.50	0.50
Naïve Bayes	0.00	0.33	0.33	0.36	0.50	0.50
C5.0	0.33	0.54	0.67	0.71	0.94	0.83
Bagged CART	0.33	0.38	0.58	0.64	0.94	0.83
Random Forest	0.33	0.54	0.67	0.71	0.94	0.83
Kappa						
GLMNET	0.00	0.10	0.53	0.55	1.00	1.00
SVM	0.00	0.00	0.00	0.00	0.00	0.00
KNN	0.00	0.00	0.00	0.15	0.00	1.00
CART	-0.50	0.00	0.00	-0.06	0.00	0.00
Naïve Bayes	-0.80	0.00	0.00	-0.09	0.00	0.00
C5.0	0.00	0.10	0.50	0.50	0.89	1.00
Bagged CART	0.00	0.00	0.25	0.41	0.90	1.00
Random Forest	0.00	0.10	0.50	0.50	0.89	1.00

(Appendix M). However, in general, it allows us to draw our main conclusion about the presence of a powerful predictive potential in the Google user search volume indicator (SVI), which, even with a small dataset and an imbalanced sample, confirms its significance.

Also, we performed the same sequence of experiments (stages 1-3 of exploring the predictive potential of the Google user search volume step) using modified versions of the targeted dataset, namely: (i) *without* considering the *time-lag* alignment, and using SVI by *search keywords* from the initial dataset (Appendix D); (ii) *without* considering the *time-lag* alignment, and using SVI by *search keywords* from the *targeted* dataset (Figure 3 and Table 8). The comparison of the validation results of the predictive models (Appendix K) additionally confirms the feasibility of using, for predicting the House Price Indices, (i) only highly correlated search keywords and (ii) SVI values with an optimal time-lag.

Thus, this step of our study made it possible (1) to confirm the assumption that the user search volume (SVI) can serve as a *sole determinant to anticipate* real estate market price shift direction; (2) to verify the *significance* of using a *cross-correlation analysis* while building targeted samples both

TABLE 11. Evaluation of the predictive ability of classification models.

Model	Accuracy	Class	Precision	Recall	F1-score
Random Forest	1	*Increase*	1	1	1
		Stable	1	1	1
		Decrease	1	1	1
Bagged CART	0.833	*Increase*	1	0.5	0.667
		Stable	0.667	1	0.800
		Decrease	1	1	1
GLMNET	0.667	*Increase*	1	0.5	0.667
		Stable	0.5	1	0.667
		Decrease	1	0.5	0.667
C5.0	0.500	*Increase*	1	0.5	0.667
		Stable	0.4	N/A	0.571
		Decrease	N/A	0	N/A

(i) for the selection of the list of relevant keywords that determine the number of input variables of the predictive model, and (ii) for identifying the presence and size of the time-lag between the investigated output and input variables; (3) to verify an *appropriate* level of accuracy in the prediction of House Price Indices changes by using machine learning algorithms based only on the Google search activity variable; that, in turn, allowed (4) to assume that the prediction accuracy of the models which are based on existing determinants of housing prices can be improved if SVI variable is included in the model.

V. DISCUSSION

Thus, in this study, we focused on analyzing the Google user search volume to assess *whether changes in the search activity with respect to information about real estate market offers and credit conditions for the purchase of real estate are informative enough to predict shift direction in real estate market prices*. The main *theoretical* contribution of our work is to demonstrate (i) how the analytical methods for studying user activity in the Internet space can potentially be used to expand the capabilities of price predictive models; and (ii) how freely available monthly information about Google users' searches allow to provide in-depth insights to enrich the generally accepted statistics on supply and demand in the real estate market. The major *contribution* of this study to the literature is the confirmation that Google user search volume can be associated as an extra determinant to anticipate the housing price shift direction with time-lag sufficient for making decisions regarding the purchase (sale) of individual property or the real estate market control.

The *methodological* contribution of our research is the composite use of known cross-correlation and ML classification approaches with the enrichment of this composition by:

(1) an approach to the multi-class classification of the HPI values, which allows predicting not only the direction (increasing or decreasing) of the price change but also its relative stability within +/−2% around 100 HPI value. This approach is a more accurate adjusting tool in contrast to the more common binary classification [32], and, in combination with the relative nature of the HPI indicator, significantly increases the forecasting results quality;

(2) an approach to *setting up input* and *output variables* characterized by different timestamps (quarters or months) with the maximum possible use of the advantages of both timestamp intervals. This made it possible (i) to more accurately identify the time-lag between the real estate price and user interest, as well as (ii) to form an *aligned* series of the target sample according to the rule: the value of user search traffic in the first month of each quarter corresponds to the House Price Index value in the next quarter;

(3) the use of *two adjacent perspectives* of user interests to form a keywords list, which makes it possible, on the one hand, to expand the range of input variables of the models, and on the other hand, to clarify the degree of influence on market prices of both (i) the user's direct interest in buying real estate and (ii) the user's indirect interest in an alternative source of funding for this type of purchase. A somewhat similar approach in terms of the use of several adjacent perspectives of user interests ('house for sale' and 'mortgage' keywords) was adopted by [35] for exploring the usefulness of Google search engine data in predicting the real estate market in Great Britain based on econometric models. In our study, however, we focused on two perspectives and then analysed two huge sets of keywords, matching the keywords with high cross-correlation coefficients to models;

(4) a phased *reduction* of the list of models input variables (SVI related to search keywords) of the predictive model using (i) an expert approach, (ii) a correlation analysis (for quarterly timestamped input and output datasets) and (iii) a cross-correlation analysis (quarterly timestamp of output data and monthly timestamp of input data). This approach provides the ability to *adjust* a targeted sample to increase the predictive model accuracy. According to [41], variable selection mechanisms, which we incorporated, are significant because GT is characterized by having some very strong predictors hidden within a huge number of weak or irrelevant terms. A similar approach, but on a much smaller scale, was tested by [32], who examined the use of Google Trends data to predict the rate of return of the WIG20 index. At the first stage, seven of the nearly 30 terms were selected, thematically related to the stock exchange and finance, that were characterized by the highest correlation coefficients. Then the possibility of using this type of data for predictive purposes was explored. In this context, two machine learning classification algorithms were used: logarithmic regression

and naive Bayes classifier. This approach produced good results, so we applied it on a much larger scale and used complex-based techniques to reduce the sample in our study.

The incremental *practical* contribution of the research is:

(1) the confirmation of the presence of a *positive correlation* between the changes in the *real estate* Google user search volume (RESVI) and the price activity with a *positive* time-lag. Concerning the Polish market, the optimal size of time-lag is equal to 4 months. Thus, we can assume that with a shift (positive or negative) in user search interest at the beginning of one quarter, we can confidently expect a correspondent shift in real estate prices at the beginning of the next quarter. As we mentioned before, prior literature (including studies especially before 2017 and outside the Scopus database) has confirmed cases of the presence of a positive correlation between user queries search traffic and price activity on the English, US and Indian housing market [39], [35], [34], [37], [2], [1]. According to [39] and [40], the prediction of house prices based on search data is very convincing, and Google data outperform existing indicators in terms of forecasting accuracy. Our results mirror those of [2] and [1] who found the existence of a positive correlation between search intensity for a term and HPI value and confirmed that this relationship is statistically significant and positively correlated with the housing price in the next quarter. On the other hand, [36] study implies that longer periods would accompany more expensive goods, which start from research online and finish in making a purchase. Given this relationship, it can be said that the 4-month delay that our study revealed is reasonable. Houses and flats are a significant expense relative to income.

(2) the understanding that Google users' search interest in information about *credit to buy real estate* (CBSVI) *correlates less* with changes in real estate prices than general information about *current offers on the real estate market*. And although there is a positive correlation between these indicators, we could less likely rely on changes in prices in the real estate market when the search volume for credit for housing changes.

(3) substantiated experimental confirmation of the main assumption of this study relating to the possibility of using Google user search volume data as an additional *significant, relevant and freely available determinant* to anticipate the price changes in the real estate market. Moreover, as evidenced by prior studies [1], [2], [35], [37], [40], Google Trends data forecast real estate prices regardless of the geographical area. This implies that commercial real estate businesses should consider incorporating this free and current dataset into their market forecasts with the possibility of using this indicator both (i) as a *sole* determinant of changes in the value of the real estate, and, (ii) in *addition* to all existing determinants of housing price. For individuals, Google Trends data are a valuable source of information for future investment decisions as well.

We are aware that our contributions have *generalization* problems and *limitations*, among which the most significant are the following:

(1) the obtained theoretical conclusions regarding the *time-lag size* in the Google user search volume and the price activity dependencies is valid for specific selected countries and will require additional research and adjustment when using other datasets and geographical locations. However, finding the presence of such a *time-lag* could be largely considered generalized, since it is confirmed both by (i) our own Granger Causality test validation results and by (ii) findings of similar studies in the field of real estate and other areas (described above).

(2) the methodological conclusion about the *significant predictive power* of SVI was made based on testing the selected machine-learning classification models and with the small dataset available for this study. We realize, when using other ML models (not tested in this study), that additional experiments and model evaluation will be required. However, the results we obtained when testing model overfitting, and which are based on small dataset (including imbalanced sample), look promising and allow us to expect our findings to be confirmed with other studies.

(3) in our study we assumed that (i) the potential housing buyers notably obtain housing information from Google search, and (ii) the search volume of keywords is related to potential housing purchase behaviour. Although such assumptions may seem problematic at first glance, they are in line with the specifics of the Polish market and prior literature. First, the buyers of the apartment here are mainly individuals, not institutions. These people rarely use the services of a real estate agent, because it is associated with a very high fee. This trend is also becoming visible in other countries, i.e., in the United States, where the share of house buyers who used the Internet to search for a house increased in 2020 to an all-time high of 97% [8]. Second, there are often no interesting offers on the agent's website whereas the individuals searching on the Internet can quickly find the property that best reflects the features sought. Third, the literature shows the intention may be the main predictor of any behaviour [75]. Thus, the purchase intention of a property supported by an online search can be regarded as the main antecedent of purchase behaviour. Studies [1], [12] are also based on these assumptions. In line with [25] many searches and associated keywords indicate a reaction to current events or predict a future event.

(4) this study does not give a direct answer to the question of whether SVI is more effective than other traditional models in the real estate market. We conducted several rough experiments with SVI as an additional variable to the predictive model that uses the quarterly sales of real estate in Poland in the period between 01.01.2011 and 31.12.2019 as a determinant of real estate price changes. The results of these experiments showed a significant increase (by 20% relative to the base) in the predictive accuracy due to the contribution of the SVI extra variable. However, based on

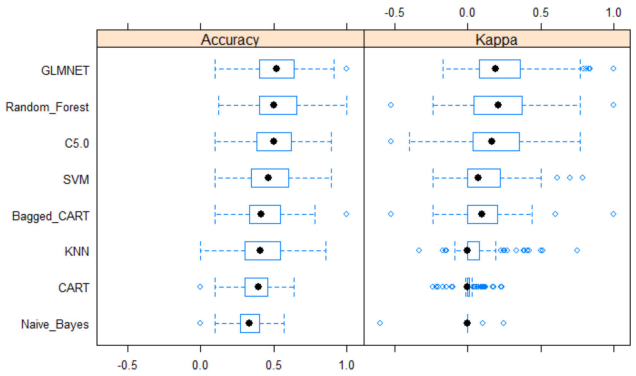

FIGURE 4. Results of validation using bootstrap algorithm.

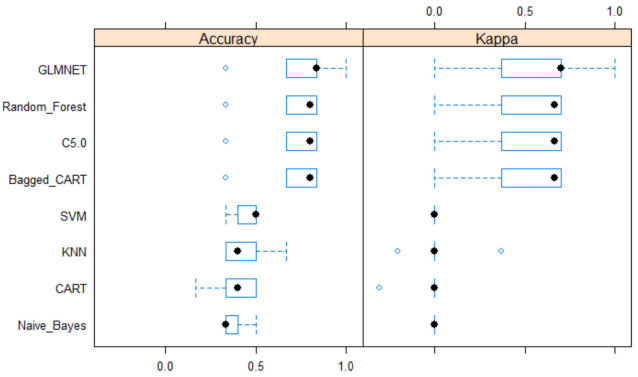

FIGURE 5. Results of validation using k-fold cross-validation algorithm.

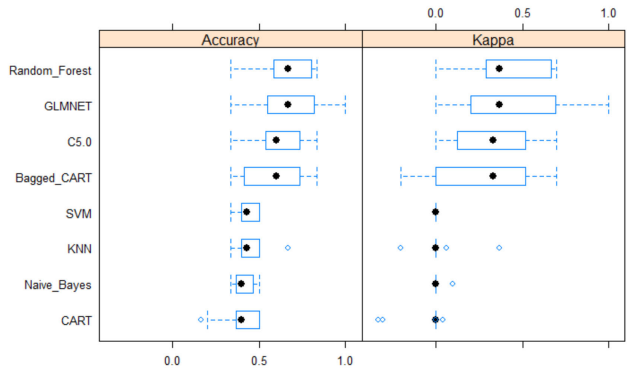

FIGURE 6. Results of validation using repeated k-fold cross-validation algorithm.

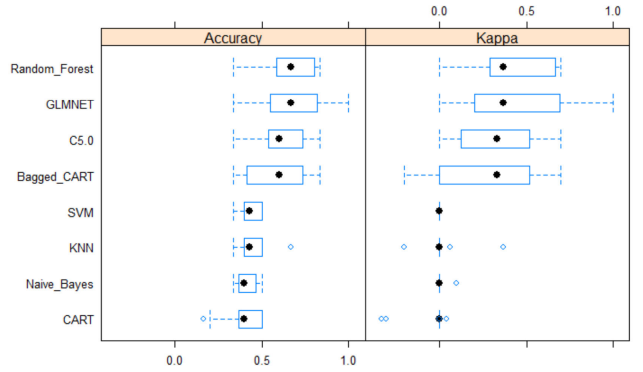

FIGURE 7. Results of validation using leave one out cross validation algorithm.

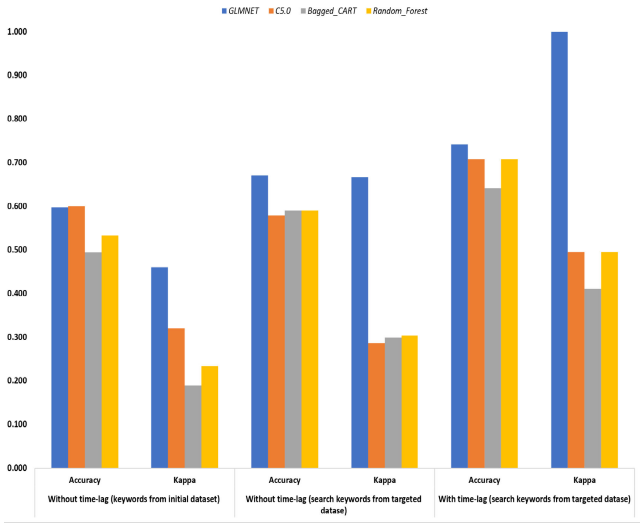

FIGURE 8. Comparison of the parameters of the predictive models' validation with different datasets.

the results of the experiments, it is still premature to draw any conclusions about the comparative effectiveness of the Google Trends data as a sole determinant in relation to the traditional ones, due to the use of, first of all, a limited number of traditional indicators. Moreover, this was not the purpose of our study. Nevertheless, we can additionally conclude that due to the facts that, *first*, SVI has demonstrated its sole predictive power of future housing price change during our main research and, *second*, taking into account that SVI data availability and variety (timestamps, regions, key-words and other options), our findings also can be *helpful* for researchers who intend to use the Google Trends data as an *extra indicator* from the real estate demand side *to improve the prediction accuracy* if it is included in the model which is based on the traditional housing prices determinants.

VI. CONCLUSION AND FUTURE WORK

The usage of Google Trends data has revolutionized many markets, and now we have demonstrated that it is significant in the housing market. Our study extended the current understanding of the research potential of the Google search engine queries volume to predict real estate price changes, taking into account users' interests in both 'real estate' and 'credit to buy real estate'. Thus, the question posed in the title of this article can be answered in the affirmative. Furthermore, we have shown that price trends in housing markets can be anticipated by the collective wisdom of online users on the web. To understand the current housing market and forecast future market trends, economists, managers from the real estate market, and investors mainly rely on housing data released from government reports. However, these

TABLE 12. Descriptive statistics of the HPI index.

Mean	105.18
Standard Error	1.30
Median	103.07
Standard Deviation	7.80
Sample Variance	60.99
Kurtosis	1.16
Skewness	1.34
Range	29.91
Minimum	97.15
Maximum	127.06
Count	36
Largest(1)	127.06
Smallest(1)	97.15
Confidence Level(95.0%)	2.64

data are released with a delay. This is a major impediment to rational decision-making. Google Trends provides many detailed reports on the volume of housing-related queries (e.g., regional analyses at the country, state and city levels). By using these data collected from GT, all entities (investors, policymakers, individuals, etc.) concerned with this topic can obtain a deeper insight into the housing market to make informed decisions. Accurate predictions about the housing market can benefit a wide spectrum of users, such as home buyers and sellers, mortgage traders, home builders, as well as individuals. Buying a house/flat is a major decision and associated with a large financial expense. To be performed rationally, it must be based on current and accurate data. Similarly, businesses that depend on the housing market can benefit from our findings. This knowledge would allow the construction industry to establish appropriate prices because the signals collected from a search can be very helpful for predicting future price trends in the real estate market.

So, the main potential channel for how the search volume can affect future housing prices is reaching to stakeholders' significant information about future housing price changes (i) *earlier* than, e.g., the official statistics and reports and (ii) with *time-lag sufficient for making decisions* regarding the purchase (sale) of individual property or the real estate market control. Such information could be used by the beneficiaries, mentioned before, for further decision making by the following rules: (i) when one discovers any changes (rise or fall) in user search interest for 'real estate' keywords at the beginning of one quarter, then may expect the change (in the same direction) or count on relative stability in real estate prices only at the beginning of the next quarter; (ii) when one sees any changes (rise or fall) in GT volume for 'credit to buy real estate' keywords at the beginning of one quarter, then may expect the change (in the same direction) or count on relative stability in real estate prices only at the end of the next quarter. As a possible scenario of reaction to the information received, for example, individual buyers or home buyers and sellers can quickly decide to buy real estate without waiting for the price increase, or, on the contrary, postpone the purchase, taking into account the anticipated decline in real estate prices in the near future.

To sum up, our findings are helpful not only for individuals but also for business practitioners, e.g., home builders, future traders and mortgage originators, to better position themselves for housing price moves. Furthermore, this knowledge may be useful in creating national policy-making for the housing market. All the named beneficiaries of our findings need them to make rational decisions regarding their needs in the real estate market. To ensure investment or business success, all of them need to be able to look ahead and plan their investments based on accurate and up-to-date data.

We are aware of the fact that Google search queries do not represent all the online housing search activities, yet in today's world, they comprise an essential part of the process for any person actively involved in a purchase or sale. Many people who want to buy a house or flat use real estate portals directly (in Poland, e.g., Morizon, DomiPorta, Sprzedajemy, Nieruchomosci-Online, etc.) and thus bypass

TABLE 13. Descriptive statistics of the "Real estate" related keywords dataset.

Keywords	"houses for sale"	"newhome"	"property for sale"	"morizon domy"	"commercial real estate"	"otodom domy na sprzedaż"	"domy na sprzedaz olx"	"dom na sprzedaż"	"mieszkanie na sprzedaż"	"domy na sprzedaż"	"mieszkania na sprzedaż"	"nieruchomości na sprzedaż"	"mieszkania na sprzedaz olx"	"mieszkania na sprzedaż olx"	"willa na sprzedaż"
Mean	118.15	73.48	128.68	86.51	81.28	116.67	97.44	122.56	92.16	110.47	132.78	159.03	94.25	64.81	70.83
Standard Error	10.03	9.90	9.56	11.69	5.84	9.89	14.96	12.87	10.89	11.38	13.81	8.56	15.29	11.41	7.42
Median	113.69	63.64	126.74	57.14	73.50	100.00	92.31	104.82	73.33	92.31	118.00	162.50	68.97	43.50	62.50
Mode	115.48	0.00	120.93	42.86	67.00	66.67	0.00	#N/A	33.33	83.33	72.00	162.50	0.00	2.00	50.00
Standard Deviation	60.19	59.41	57.33	70.15	35.05	59.36	89.78	77.22	65.33	68.27	82.84	51.37	91.71	68.48	44.52
Kurtosis	-0.77	-1.11	-0.76	-0.55	0.57	-0.63	-1.57	-1.15	-0.27	-0.81	-1.28	0.02	-1.33	-0.58	2.76
Skewness	0.32	0.21	-0.05	0.79	0.80	0.37	0.24	0.34	0.83	0.49	0.22	0.28	0.48	0.83	1.29
Range	219.05	181.82	220.93	257.14	150.00	233.33	246.15	256.63	224.44	246.15	280.00	212.50	262.07	212.00	225.00
Minimum	10.71	0.00	15.12	0.00	27.00	0.00	0.00	16.87	13.33	14.10	12.00	75.00	0.00	0.00	0.00
Maximum	229.76	181.82	236.05	257.14	177.00	233.33	246.15	273.49	237.78	260.26	292.00	287.50	262.07	212.00	225.00
Largest(1)	229.76	181.82	236.05	257.14	177.00	233.33	246.15	273.49	237.78	260.26	292.00	287.50	262.07	212.00	225.00
Smallest(1)	10.71	0.00	15.12	0.00	27.00	0.00	0.00	16.87	13.33	14.10	12.00	75.00	0.00	0.00	0.00
Confidence Level	20.37	20.10	19.40	23.73	11.86	20.09	30.38	26.13	22.10	23.10	28.03	17.38	31.03	23.17	15.06

TABLE 14. Descriptive statistics of "Credit to buy real estate" related keywords dataset.

Keywords	"kredyt na mieszkanie"	"kredyt na dom"	"wkład własny mieszkanie"
Mean	110.68	159.98	95.30
Standard Error	7.29	8.12	8.07
Median	109.15	153.70	94.44
Mode	50.70	200.00	18.52
Standard Deviation	43.71	48.75	48.42
Kurtosis	-1.07	-0.78	-0.61
Skewness	0.01	0.10	0.22
Range	152.11	203.70	171.60
Minimum	38.03	59.26	18.52
Maximum	190.14	262.96	190.12
Largest(1)	190.14	262.96	190.12
Smallest(1)	38.03	59.26	18.52
Confidence Level(95.0%)	14.79	16.49	16.38

TABLE 15. Granger causality test for "Real estate" keywords dataset.

Lag (months)	Direction	F-statistic	Pr(>F)	Significance
1	REUI→HPI	37.725	7.20E-07	***
1	HPI→REUI	0.160	0.69220	-
2	REUI→HPI	17.998	8.28E-06	***
2	HPI→REUI	0.894	0.00100	-
3	REUI→HPI	11.169	6.82E-05	***
3	HPI→REUI	2.453	0.08576	-
4	REUI→HPI	8.258	2.78E-04	***
4	HPI→REUI	3.315	0.02772	*
5	REUI→HPI	3.7612	0.01457	*
5	HPI→REUI	0.100	0.10000	-
6	REUI→HPI	2.505	0.06368	-
6	HPI→REUI	1.361	0.28560	-

TABLE 16. Granger causality test for "Credit to buy real estate" keywords dataset.

Lag (months)	Direction	F-statistic	Pr(>F)	Significance
1	CBSVI →HPI	47.254	8.87E-08	***
1	HPI→ CBSVI	2.3037	0.13890	-
2	CBSVI →HPI	21.942	1.57E-06	***
2	HPI→ CBSVI	1.101	0.34590	-
3	CBSVI →HPI	13.002	2.22E-05	***
3	HPI→ CBSVI	0.468	0.70740	-
4	CBSVI →HPI	9.592	1.03E-04	***
4	HPI→ CBSVI	9.592	1.03E-04	***
4	CBSVI →HPI	0.423	0.79070	-
5	HPI→ CBSVI	5.5672	0.03937	*
5	CBSVI →HPI	1.419	0.2727	-
6	HPI→ CBSVI	4.7618	0.236	-

the search engine. However, we believe the same as [34] that, despite missing some possible segments of the population for the above-mentioned reason, we can still predict the change in housing price using only an online search captured by Google SVI. This clearly shows the power of online queries in forecasting economic trends, possible for usage in many fields. Therefore, data collected from Google Trends (SVI) could prove to be one of the most powerful tools for helping clients, businesses and government officials make precise future economic predictions and thus, optimal decisions.

In the future, the methodology developed for this research will be tested in areas other than the real estate market as well as using data about the users' activity in other Internet resources, for example, Twitter, Reddit. This will create the next stage of its improvement and development.

APPENDIX

A. APPENDIX 1

Real Estate Keywords https://github.com/TextAnalitics/Prices-Prediction-for-Real-Estate-Market/blob/main/Appendix%201.%20Real%20Estate%20Keywords.pdf

B. APPENDIX 2

Keywords concerning buying various types of real estate https://github.com/TextAnalitics/Prices-Prediction-for-Real-Estate-Market/blob/main/Appendix%202.%20Keywords%20concerning%20buy%20various%20types%20of%20real%20estate.pdf

C. APPENDIX 3

Keywords possibility of using credit to buy real estate https://github.com/TextAnalitics/Prices-Prediction-for-Real-Estate-Market/blob/main/Appendix%203.%20Keywords%20possibility%20of%20using%20a%20credit%20to%20buy%20real%20estate.pdf

TABLE 17. Checking ML models overfitting.

Algorithms	Prediction accuracy of the training dataset			Prediction accuracy of the test dataset
	Min.	Mean	Max	
GLMNET	0.33	0.74	1.00	0.67
C5.0	0.33	0.71	0.83	0.50
Bagged CART	**0.33**	**0.64**	**0.83**	**0.83**
Random Forest	**0.33**	**0.71**	**0.83**	**1.00**

TABLE 18. Structure of the predictive model's training and test samples.

Class labels	Training Subsample	Test Subsample	Total by Classes
Increase	12	2	14
Stable	13	2	15
Decrease	4	2	6
Total by Sample	29	6	35

D. APPENDIX 4

Selected keywords https://github.com/TextAnalitics/Prices-Prediction-for-Real-Estate-Market/blob/main/Appendix%204.%20Selected%20keywords.pdf

E. APPENDIX 5

Dataset https://github.com/TextAnalitics/Prices-Prediction-for-Real-Estate-Market/blob/main/Appendix%205.%20Full%20Dataset.xlsx

F. APPENDIX 6

See Table 12.

G. APPENDIX 7

See Table 13.

H. APPENDIX 8

See Table 14.

I. APPENDIX 9

Granger causality test. See Table 15. The directions of causality in each test have the following interpretation: for example, REUI → HPI means that the null hypothesis is "*REUI does not Granger-cause HPI*". The condition for rejecting the null hypothesis is the condition under which $p <= 0.05$.

1) APPENDIX 9

See Table 16.

2) APPENDIX 9

See Table 17.

TABLE 19. Summary of the evaluation of predictive algorithms.

Algorithms	Min.	1st Quartile	Median	Mean	3rd Quartile	Max
Accuracy						
GLMNET	0.50	0.67	0.80	0.76	0.83	1.00
SVM	0.33	0.40	0.50	0.45	0.50	0.50
KNN	0.33	0.40	0.50	0.51	0.67	0.67
CART	0.33	0.40	0.50	0.45	0.50	0.50
Naïve Bayes	0.33	0.40	0.50	0.45	0.50	0.50
C5.0	0.50	0.67	0.80	0.73	0.83	0.83
Bagged CART	0.50	0.50	0.67	0.66	0.80	0.83
Random Forest	0.50	0.50	0.67	0.66	0.80	0.83
Kappa						
GLMNET	0.22	0.37	0.67	0.59	0.70	1.00
SVM	0.00	0.00	0.00	0.00	0.00	0.00
KNN	-0.09	0.12	0.28	0.21	0.37	0.37
CART	0.00	0.00	0.00	0.00	0.00	0.00
Naive Bayes	0.00	0.00	0.00	0.00	0.00	0.00
C5.0	0.22	0.37	0.67	0.53	0.70	0.70
Bagged CART	0.00	0.22	0.45	0.42	0.67	0.75
Random Forest	0.00	0.22	0.45	0.41	0.67	0.70

TABLE 20. Evaluation of the predictive ability of classification models.

Model	Accuracy	Class	Precision	Recall	F1-score
Random Forest	0.666667	Increase	1	0.5	0.6667
		Stable	0.5	1	0.6667
		Decrease	1	0.5	0.6667
GLMNET	0.666667	Increase	1	0.5	0.6667
		Stable	0.5	1	0.6667
		Decrease	1	0.5	0.6667
Bagged CART	*0.5*	*Increase*	*1*	*0.5*	*0.6667*
		Stable	*0.4*	*N/A*	*0.5714*
		Decrease	*N/A*	*0*	*N/A*
C5.0	*0.33*	*Increase*	*N/A*	*0*	*N/A*
		Stable	*0.33*	*1*	*0.5*
		Decrease	*N/A*	*0*	*N/A*

J. APPENDIX 10

See Table 18. Comparison of the predictive models using different validation algorithms.

Bootstrap validation algorithm involves taking random samples from the dataset (with re-selection) against which to evaluate the model. In aggregate, the results indicate the variance of the model's performance.

The k-fold cross-validation method involves splitting the dataset into k-subsets. Each subset is held out while the model is trained on all other subsets. This process is completed until accuracy is determined for each instance in the dataset, and an overall accuracy estimate is provided.

The process of splitting the data into k-folds can be repeated a number of times, this is called Repeated k-fold cross-validation. The final model accuracy is taken as the mean from the number of repeats.

In Leave One Out Cross Validation (LOOCV), a data instance is left out and a model is constructed on all other data instances in the training set. This is repeated for all data instances.

K. APPENDIX 11

See Table 19.

L. APPENDIX 12

Predictive value of Google users interest data analysis with imbalanced training and test subsamples.

M. APPENDIX 13

See Table 20.

REFERENCES

[1] M. Venkataraman, V. Panchapagesan, and E. Jalan, "Does Internet search intensity predict house prices in emerging markets? A case of India," *Property Manage.*, vol. 36, no. 1, pp. 103–118, Feb. 2018, doi: 10.1108/PM-01-2017-0003.

[2] E. Beracha and M. B. Wintoki, "Forecasting residential real estate price changes from online search activity," *J. Real Estate Res.*, vol. 35, no. 3, pp. 283–312, Jun. 2013, doi: 10.5555/rees.36.3.6417244666x72788.

[3] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine Query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, Feb. 2009.

[4] I. Bordino, S. Battiston, G. Caldarelli, M. Cristelli, A. Ukkonen, and I. Weber, "Web search queries can predict stock market volumes," *PLoS ONE*, vol. 7, no. 7, Jul. 2012, Art. no. e40014, doi: 10.1371/journal.pone.0040014.

[5] S. Mihaela, "Improving unemployment rate forecasts at regional level in romania using Google trends," *Technol. Forecasting Social Change*, vol. 155, Jun. 2020, Art. no. 120026.

[6] F. Takeda and T. Wakao, "Google search intensity and its relationship with returns and trading volume of Japanese stocks," *Pacific-Basin Finance J.*, vol. 27, pp. 1–18, Oct. 2014. [Online]. Available: https://ssrn.com/abstract=2332495, doi: 10.2139/ssrn.23324952018.

[7] R. Tao, X. Zhang, and L. Zhao, "Forecasting crude oil prices based on an Internet search driven model," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 4156–4161, doi: 10.1109/bigdata.2018.8622152.

[8] (2020). *National Association of Realtors Report*. [Online]. Available: https://www.nar.realtor/research-and-statistics/research-reports

[9] J. Surowiecki, *The Wisdom of Crowds*. New York, NY, USA: Anchor Book, 2005.

[10] H. Hong, Q. Ye, Q. Du, G. A. Wang, and W. Fan, "Crowd characteristics and crowd wisdom: Evidence from an online investment community," *J. Assoc. Inf. Sci. Technol.*, vol. 71, no. 4, pp. 423–435, Apr. 2020, doi: 10.1002/asi.24255.

[11] H. Choi and H. Varian, *Predicting Initial Claims for Unemployment Benefits*. Menlo Park, CA, USA: Google, 2012.

[12] H. Choi and H. Varian, "Predicting the present with Google trends," *Econ. Rec.*, vol. 88, no. 1, pp. 2–9, 2012.

[13] S.-P. Jun, H. S. Yoo, and S. Choi, "Ten years of research change using Google trends: From the perspective of big data utilizations and applications," *Technol. Forecasting Social Change*, vol. 130, pp. 69–87, May 2018.

[14] L. Frauenfeld, D. Nann, Z. Sulyok, Y.-S. Feng, and M. Sulyok, "Forecasting tuberculosis using diabetes-related Google trends data," *Pathogens Global Health*, vol. 114, no. 5, pp. 236–241, Jul. 2020.

[15] S. B. Choi and I. Ahn, "Forecasting seasonal influenza-like illness in South Korea after 2 and 30 weeks using Google Trends and influenza data from Argentina," *PLoS ONE*, vol. 15, no. 7, Jul. 2020, Art. no. e0233855.

[16] W. Anggraeni and L. Aristiani, "Using Google trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia," in *Proc. Int. Conf. Inf. Commun. Technol. Syst. (ICTS)*, 2016, pp. 114–118.

[17] J. M. Barros, R. Melia, K. Francis, J. Bogue, M. O'sullivan, K. Young, R. A. Bernert, D. Rebholz-Schuhmann, and J. Duggan, "The validity of Google trends search volumes for behavioral forecasting of national suicide rates in Ireland," *Int. J. Environ. Res. Public Health*, vol. 16, no. 17, Sep. 2019, Art. no. 3201.

[18] Y. Chai, H. Luo, Q. Zhang, Q. Cheng, C. S. M. Lui, and P. S. F. Yip, "Developing an early warning system of suicide using Google trends and media reporting," *J. Affect. Disorders*, vol. 255, pp. 41–49, Aug. 2019.

[19] U. Gunter, I. Önder, and S. Gindl, "Exploring the predictive ability of LIKES of posts on the facebook pages of four major city DMOs in austria," *Tourism Econ.*, vol. 25, no. 3, pp. 375–401, May 2019.

[20] E. Silva, H. Hassani, D. Madsen, and L. Gee, "Googling fashion: Forecasting fashion consumer behaviour using Google trends," *Social Sci.*, vol. 8, no. 4, p. 111, Apr. 2019.

[21] M. Dilmaghani, "Workopolis or the pirate bay: What does Google trends say about the unemployment rate?" *J. Econ. Stud.*, vol. 46, no. 2, pp. 422–445, Mar. 2019.

[22] Š. Lyócsa, E. Baumöhl, T. Výrost, and P. Molnár, "Fear of the coronavirus and the stock markets," *Finance Res. Lett.*, vol. 36, Oct. 2020, Art. no. 101735.

[23] A. A. Salisu, A. E. Ogbonna, and A. Adewuyi, "Google trends and the predictability of precious metals," *Resour. Policy*, vol. 65, Mar. 2020, Art. no. 101542.

[24] M. Qadan and H. Nama, "Investor sentiment and the price of oil," *Energy Econ.*, vol. 69, pp. 42–58, Jan. 2018.

[25] K. Wołk, "Advanced social media sentiment analysis for short-term cryptocurrency price prediction," *Expert Syst.*, vol. 37, no. 2, Apr. 2020, Art. no. e12493.

[26] M. Seo, S. Lee, and G. Kim, "Forecasting the volatility of stock market index using the hybrid models with Google domestic trends," *Fluctuation Noise Lett.*, vol. 18, no. 1, Mar. 2019, Art. no. 1950006.

[27] H. Hu, L. Tang, S. Zhang, and H. Wang, "Predicting the direction of stock markets using optimized neural networks with Google trends," *Neurocomputing*, vol. 285, pp. 188–195, Apr. 2018.

[28] Y. Wang and H. Wang, "Using networks and partial differential equations to forecast bitcoin price movement," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 30, no. 7, Jul. 2020, Art. no. 073127.

[29] J. Parker, C. Cuthbertson, S. Loveridge, M. Skidmore, and W. Dyar, "Forecasting state-level premature deaths from alcohol, drugs, and suicides using Google trends data," *J. Affect. Disorders*, vol. 213, pp. 9–15, Apr. 2017.

[30] S. Emili, A. Gardini, and E. Foscolo, "High spatial and temporal detail in timely prediction of tourism demand," *Int. J. Tourism Res.*, vol. 22, no. 4, pp. 451–463, Jul. 2020.

[31] P. P. Schneider, C. J. A. W. van Gool, P. Spreeuwenberg, M. Hooiveld, G. A. Donker, D. J. Barnett, and J. Paget, "Using Web search queries to monitor influenza-like illness: An exploratory retrospective analysis, Netherlands, 2017/18 influenza season," *Eurosurveillance*, vol. 25, no. 21, 2020, Art. no. 1900221.

[32] E. Niedzielska, "Using Google trends to predict the rate of return of WIG20 index," *Econ. Century*, vol. 3, no. 19, pp. 82–97, 2018.

[33] I. Cara, J. P. Paardekooper, and T. Helmond, "The potential of applying machine learning for predicting cut-in behaviour of surrounding traffic for truck-platooning safety," in *Proc. 25th Int. Tech. Conf. Enhanced Saf. Vehicles (ESV) Nat. Highway Traffic Saf. Admin.*, 2017, pp. 3–8.

[34] L. Wu and E. Brynjolfsson, "The future of prediction: How Google searches foreshadow housing prices and sales," in *Proc. Econ. Anal. Digit. Economy*, Apr. 2009, pp. 89–118.
[35] G. Bulczak, "Zastosowanie Google Trends w prognozowaniu zmian na rynku nieruchomości," *J. Manage. Finance*, vol. 12, no. 4, pp. 79–90, 2014.
[36] G. J. Stigler, "The economics of information," *J. Political Econo.*, vol. 69, p. 3, pp. 213–225, 1961.
[37] R. Hohenstatt, M. Kasbauer, and W. Schaffers, "'Geco' and its potential for real estate research: Evidence from the US housing market" *J. Real Estate Res.*, vol. 33, no. 4, pp. 471–506, 2011.
[38] R. Kulkarni, K. E. Haynes, R. R. Stough, and J. H. P. Paelinck, "Forecasting housing prices with Google econometrics," GMU School Public Policy, Hopewell, VA, USA, Res. Paper 2009-10, 2009.
[39] N. McLaren and R. Shanbhogue, "Using Internet search data as economic indicators," *Quart. Bull.*, vol. 2, pp. 134–140, Oct. 2011.
[40] M. Alexander Dietzel, N. Braun, and W. Schäfers, "Sentiment-based commercial real estate forecasting with Google search volume data," *J. Property Investment Finance*, vol. 32, no. 6, pp. 540–569, Aug. 2014, doi: 10.1108/JPIF-01-2014-0004.
[41] D. Borup, E. Christian, and M. Schütte, "In search of a job: Forecasting employment growth using Google trends," *J. Bus. Econ. Statist.*, Jul. 2020, doi: 10.1080/07350015.2020.1791133.
[42] G. K. Webb, "Internet search statistics as a source of business intelligence: Searches on foreclosure as an estimate of actual home foreclosures," *Issues Inf. Syst.*, vol. 10, no. 2, p. 82, 2009.
[43] X. Wang, J. Wen, Y. Zhang, and Y. Wang, "Real estate price forecasting based on SVM optimized by PSO," *Optik*, vol. 125, no. 3, pp. 1439–1443, Feb. 2014.
[44] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 259–268, Jan. 2015, doi: 10.1016/j.eswa.2014.07.040.
[45] B. Park and J. K. Bae, "Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 2928–2934, Apr. 2015, doi: 10.1016/j.eswa.2014.11.040.
[46] T. Harris, "Credit scoring using the clustered support vector machine," *Expert Syst. Appl.*, vol. 42, no. 2, pp. 741–750, Feb. 2015, doi: 10.1016/j.eswa.2014.08.029.
[47] G. Daqi and Z. Tao, "Support vector machine classifiers using RBF kernels with clustering-based centers and widths," in *Proc. Int. Joint Conf. Neural Netw.*, Orlando, FL, USA, 2007, pp. 2971–2976, doi: 10.1109/IJCNN.2007.4371433.
[48] T. Mohd, N. S. Jamil, N. Johari, L. Abdullah, and S. Masrom, "An overview of real estate modelling techniques for house price prediction," in *Charting a Sustain. Future ASEAN Business Social Sciences*, N. Kaur and M. Ahmad, Eds. Singapore: Springer, 2020, doi: 10.1007/978-981-15-3859-9_28.
[49] M. Shahhosseini, G. Hu, and H. Pham, "Optimizing ensemble weights for machine learning models: A case study for housing price prediction," in *Smart Service Systems, Operations Management, and Analytics*, Y. H. Qiu and R. Chen, Eds. Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-30967-1_9.
[50] L. Mrsic, H. Jerkovic, and M. Balkovic, "Real estate market price prediction framework based on public data sources with case study from croatia," in *Intelligent Information and Database Systems* (Communications in Computer and Information Science), vol. 1178, P. Sitek, M. Pietranik, M. Krätkiewicz, and C. Srinilta, Eds. Singapore: Springer, 2020, doi: 10.1007/978-981-15-3380-8_2.
[51] M. F. Mukhlishin, R. Saputra, and A. Wibowo, "Predicting house sale price using fuzzy logic, artificial neural network and K-nearest neighbor," in *Proc. 1st Int. Conf. Informat. Comput. Sci. (ICICoS)*, Semarang, IN, USA, 2017, pp. 171–176, 2017, doi: 10.1109/ICICOS.2017.8276357.
[52] H. Guo, "A new technique to predict fly-rock in bench blasting based on an ensemble of support vector regression and GLMNET," *Eng. Comput.*, vol. 37, pp. 421–435, Oct. 2021, doi: 10.1007/s00366-019-00833-x.
[53] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, p. 1, 2010.
[54] T. Hastie and J. Qian. (2014). *Glmnet Vignette*. Accessed: Jun. 9, 2016. [Online]. Available: https://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf
[55] J. Manasa, R. Gupta, and N. S. Narahari, "Machine learning based predicting house prices using regression techniques," in *Proc. 2nd Int. Conf. Innov. Mech. Ind. Appl. (ICIMIA)*, Bengaluru, India, 2020, pp. 624–630, doi: 10.1109/ICIMIA48430.2020.9074952.
[56] A. Chaturvedi, "Parameterized Comparison of Regularized Regression Models to Develop Models for Real Estate," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1099, Feb. 2021, Art. no. 012016. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012016/meta
[57] S. Xiong, Q. Sun, and A. Zhou, "Improve the house price prediction accuracy with a stacked generalization ensemble model," in *Internet Vehicles Technology Services Toward Smart Cities* (Lecture Notes in Computer Science), vol. 11894, C. Hsu, S. Kallel, K. C. Lan, and Z. Zheng, Eds. Cham, Switzerland: Springer, 2020, pp. 382–385, doi: 10.1007/978-3-030-38651-1_32.
[58] A. J. Van der Kooij, "Prediction accuracy and stability of regression with optimal scaling transformations," Ph.D. dissertation, Dept. Educ. Child Stud., Fac. Social Behavioural Sci., Leiden Univ., Leiden, The Netherlands, 2007. [Online]. Available: https://openaccess.leidenuniv.nl/handle/1887/12096
[59] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer-Verlag, 2003, doi: 10.1007/978-0-387-84858-7.
[60] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc., B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005.
[61] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., Ser. B Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
[62] A. E. Hoerl and R. W. Kennard, "Ridge regression: Applications to nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 69–82, 1970.
[63] J. M. Martínez-Martínez, P. Escandell-Montero, E. Soria-Olivas, J. D. Martín-Guerrero, R. Magdalena-Benedito, and J. Gómez-Sanchis, "Regularized extreme learning machine for regression problems," *Neurocomputing*, vol. 74, no. 17, pp. 3716–3721, Oct. 2011, doi: 10.1016/j.neucom.2011.06.013.
[64] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Oct. 1995, doi: 10.1007/BF00994018.
[65] P. Durganjali and M. V. Pujitha, "House resale price prediction using classification algorithms," in *Proc. Int. Conf. Smart Struct. Syst. (ICSSS)*, Chennai, India, 2019, pp. 1–4, doi: 10.1109/ICSSS.2019.8882842.
[66] A. Babakura, M. N. Sulaiman, and M. A. Yusuf, "Improved method of classification algorithms for crime prediction," in *Proc. Int. Symp. Biometrics Secur. Technol. (ISBAST)*, Aug. 2014, pp. 250–255, doi: 10.1109/ISBAST.2014.7013130.
[67] R. Sawant, Y. Jangid, T. Tiwari, S. Jain, and A. Gupta, "Comprehensive analysis of housing price prediction in pune using multi-featured random forest approach," in *Proc. 4th Int. Conf. Comput. Commun. Control Autom. (ICCUBEA)*, Aug. 2018, pp. 1–5, doi: 10.1109/ICCUBEA.2018.8697402.
[68] K. Yoo, H. Yoo, J. M. Lee, S. K. Shukla, and J. Park, "Classification and regression tree approach for prediction of potential hazards of urban airborne bacteria during Asian dust events," *Sci. Rep.*, vol. 8, no. 1, Dec. 2018, Art. no. 11823, doi: 10.1038/s41598-018-29796-7.
[69] M. A. Razi and K. Athappilly, "A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models," *Expert Syst. Appl.*, vol. 29, no. 1, pp. 65–74, 2005.
[70] L. Breiman, *Classification and Regression Trees*, 1st ed. Evanston, IL, USA: Routledge, 1984, doi: 10.1201/9781315139470.
[71] S. Pang and J.-Z. Gong, "C5.0 classification algorithm and application on individual credit evaluation of banks," *Syst. Eng.-Theory Pract.*, vol. 29, no. 12, pp. 94–104, 2009, doi: 10.1016/S1874-8651(10)60092-0.
[72] G.-Z. Fan, S. E. Ong, and H. C. Koh, "Determinants of house price: A decision tree approach," *Urban Stud.*, vol. 43, no. 12, pp. 2301–2315, Nov. 2006, doi: 10.1080/00420980600990928.
[73] X. N. Bui, "A lasso and elastic-net regularized generalized linear model for predicting blast-induced air over-pressure in open-pit mines," *Inżynieria Mineralna*, vol. 21, p. 41, Oct. 2019.
[74] V. Plakandaras, R. Gupta, P. Gogas, and T. Papadimitriou, "Forecasting the US real house price index," *Econ. Model.*, vol. 45, pp. 259–267., Feb. 2015, doi: 10.2139/ssrn.2431627.
[75] M. Fishbein and I. Ajzen, "Belief, attitude, intention, and behavior: An introduction to theory and research," *Philosophy Rhetoric*, vol. 10, no. 2, pp. 130–132, 1977.

[76] V. Kotu, B. Deshpande, D. M. Process, E. V. Kotu, and B. Deshpande, *Predictive Analytics and Data Mining*. Burlington, AM, USA: Morgan Kaufmann, 2015, pp. 17–36, doi: 10.1016/B978-0-12-801460-8.00002-1.
[77] M. Balamurugan and S. Kannan, "Performance analysis of cart and C5.0 using sampling techniques," in *Proc. IEEE Int. Conf. Adv. Comput. Appl. (ICACA)*, Oct. 2016, pp. 72–75, doi: 10.1109/ICACA.2016.7887926.
[78] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, Sep. 2007, doi: 10.1016/j.energy.2006.11.010.
[79] R. Quinlan, "Data mining tools see5 and C5.0," in *Proc. Rulequest Res.*, Oct. 2008, pp. 1997–2004.
[80] V. Business and V. Business, "Application of hybrid methodology to predict housing loan defaults in India," *J. Int. Manage. Stud.*, vol. 15, no. 3, pp. 43–50, Dec. 2015.
[81] K. Mishra and R. Rani, "Churn prediction in telecommunication using machine learning," in *Proc. Int. Conf. Energy, Commun., Data Anal.s Soft Comput. (ICECDS)*, Chennai, India, Aug. 2017, pp. 2252–2257, doi: 10.1109/ICECDS.2017.8389853.
[82] S. N. Brohi, T. R. Pillai, S. Kaur, H. Kaur, S. Sukumaran, and D. Asirvatham, "Accuracy Comparison of Machine Learning Algorithms for Predictive Analytics in Higher Education," in *Emerging Technologies in Computing* (Lecture Notes of the Institute for Computer Sciences), vol. 285, M. Miraz, P. Excell, A. Ware, S. Soomro, and A. Ali, Eds. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-23943-5_19.
[83] I. Matijosaitiene, A. McDowald, and V. Juneja, "Predicting safe parking spaces: A machine learning approach to geospatial urban and crime data," *Sustainability*, vol. 11, no. 10, p. 2848, May 2019, doi: 10.3390/su11102848.
[84] P. Grover. (2017). *Gradient Boosting from Scratch*. Accessed: Jul. 20, 2018. [Online]. Available: https://medium.com/mlreview/gradientboosting-from-scratch-1e317ae4587d
[85] H. Pham and S. Olafsson, "Bagged ensembles with tunable parameters," *Comput. Intell.*, vol. 35, no. 1, pp. 184–203, 2019.
[86] A. Varma, A. Sarma, S. Doshi, and R. Nair, "House price prediction using machine learning and neural networks," in *Proc. 2nd Int. Conf. Inventive Commun. Comput. Technol. (ICICCT)*, Coimbatore, India, Apr. 2018, pp. 1936–1939, doi: 10.1109/ICICCT.2018.8473231.
[87] M. Áeh, M. Kilibarda, A. Lisec, and B. Bajat, "Estimating the performance of random forest versus multiple regression for predicting prices of the apartments," *ISPRS Int. J. Geo-Inf.*, vol. 7, p. 168, Oct. 2018, doi: 10.3390/ijgi7050168.
[88] S. Jamil, T. Mohd, S. Masrom, and N. Ab Rahim, "Machine learning price prediction on green building prices," in *Proc. IEEE Symp. Ind. Electron. Appl. (ISIEA)*, Kuala Lumpur, Malaysia, 2020, pp. 1–6, doi: 10.1109/ISIEA49364.2020.9188114.
[89] M. Puliga, G. Caldarelli, and S. Battiston, "Credit default swaps networks and systemic risk," *Sci. Rep.*, vol. 4, no. 1, May 2015, Art. no. 6822, doi: 10.1038/srep06822.
[90] Ü. Agbulut, A. E. Gürel, and Y. Biçen, "Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison," *Renew. Sustain. Energy Rev.*, vol. 135, pp. 1321–1364, Oct. 2021, doi: 10.1016/j.rser.2020.110114.
[91] E. A. Freeman, G. G. Moisen, J. W. Coulston, and B. T. Wilson, "Random forests and stochastic gradient boosting for predicting tree canopy cover: Comparing tuning processes and model performance," *Can. J. Forest Res.*, vol. 46, no. 3, pp. 323–339, Mar. 2016, doi: 10.1139/cjfr-2014-0562.

NINA RIZUN received the Ph.D. degree in technical sciences from the Faculty of Enterprise Economy and Production Organization, National Mining Academy, Dnipropetrovsk, Ukraine. She is currently an Assistant Professor with the Department of Informatics in Management, Faculty of Management and Economics, Gdańsk University of Technology. Her main research interests include the application of big data and text analytics methods for service quality evaluation, decision-making logic discovery, and business sentiment analysis. Most recent publications are focused on structural and temporal topic models of on service quality feedbacks, assessing business process complexity based on textual data, and mapping determinants of unintended negative consequences of disruptive technologies use in smart cities.

ANNA BAJ-ROGOWSKA received the Ph.D. degree in economic sciences in the field of management from the University of Gdansk, Poland. She is currently an Assistant Professor with the Department of Informatics in Management, Gdańsk University of Technology, Poland. Her scientific research interest includes the intersection of management and IT. She deals primarily with research regarding the technological and economic aspects of IT in organizations as well as innovations and IT-based organizational creativity. Her current research interests focus on issues of extracting information from web content using text mining algorithms and sentiment analysis techniques. She examines the impact of knowledge acquired from textual data on the contemporary business.