Journal of Universal Computer Science, vol. 28, no. 5 (2022), 540-562
submitted: 3/9/2021, accepted: 21/2/2022, appeared: 28/5/2022

A methodological framework for the integration of machine learning algorithms into agent-based simulation models

Nikola Zornić
(University of Belgrade – Faculty of Organisational Sciences, Belgrade, Serbia
https://orcid.org/0000-0002-3597-0627, nikola.zornic@fon.bg.ac.rs)

Aleksandar Marković
(University of Belgrade – Faculty of Organisational Sciences, Belgrade, Serbia
https://orcid.org/0000-0002-3976-2788, aleksandar.markovic@fon.bg.ac.rs)

Abstract: Traditionally, agent-based modelling and simulation relied on using utility function in agents' decision-making process. Some drawbacks in this process are identified, and a potential remedy to the issue is proposed. This paper introduces a methodological framework for building a hybrid agent-based model that aims to overcome some of the elaborated problems related to the usage of a utility function. In the proposed approach, a machine learning algorithm substitutes the utility function, thus providing a possibility to use various algorithms. The proposed methodological framework has been applied to a case study of churn in a telecommunications company. Three models have been created and used for simulation experiments, two using the proposed methodology and one using utility function. The pros and cons of different approaches are identified and discussed.

Keywords: agent-based, simulation, machine learning, framework, churn
Categories: I.2.1, I.6.1, I.6.3, I.6.5
DOI: 10.3897/jucs.73924

1 Introduction

Modelling and simulation have a basis in scientific fields such as mathematics, operational research, computer science, statistics, physics and engineering. They cover a wide range of methods and techniques, which enable their application in various fields, such as business, health, ecology, medicine, logistics, manufacturing, service delivery, supply chains, and many others [Ali et al., 2019, Jahangirian et al., 2010, Mourtzis et al., 2014]. According to Taylor [Taylor, 2014], the origin of modelling and simulation as a scientific discipline can be taken as the origin of the Monte Carlo simulation, i.e. Buffon's needle experiment in 1777. The basis for modern modelling and simulation was laid by the progress made during World War II [Goldsman et al., 2010]. During the analysis of industrial systems using modelling and simulation, two fundamental simulation techniques were defined [Taylor, 2014]: discrete-event simulation (DES) and system dynamics (SD). Both methods aim to analyse the behaviour of the system over time. In DES, the system is presented as a set of discrete changes of state in time (events) organised in activities [Tocher and Owen, 1960].

Independent of this approach, Forrester introduced SD [Forrester, 1961], where the system presents itself as a set of cyclic independent subsystems or feedback circuits. The application of causal-loop diagrams describes how parts of the system affect each other and feedback loops are observed. By applying more formal, stock and flow diagrams, the model can be quantified and simulation experiments performed.

The complexity of dynamic systems imposes the need to build models in order to study their functioning. Moreover, analysing the way the system works using its model is an important source of information to increase understanding of the system. DES and SD are used for modelling and simulation of a large number of systems. However, for the needs of modelling and simulation of complex adaptive systems, where the behaviour of the system depends on the interaction of a large number of system components, a special simulation technique was developed in the late 1980s – agent-based modelling and simulation (ABMS).

ABMS is most often used to model complex systems in which it is necessary to monitor the interactions of its key elements. To represent the components of the system, this methodology uses autonomous agents that have their own characteristics, decision rules and the ability to interact with other agents in the system, as well as with the environment. In addition, agents have the ability to adapt their behaviour to the circumstances arising from mutual interactions [Abbott, 2006, Prokopenko et al., 2009].

When building an agent-based simulation model, modellers typically use a large number of input and output variables, which influence the agents' behaviour. This influence is modelled using the utility function. It quantifies the values in the model, which are used for decision making by agents, although the nature of human decision-making behaviour can be unreliable [Geng et al., 2020]. The aim of this paper is to overcome these problems by developing a methodological framework for the integration of machine learning algorithms in place of a utility function. The machine learning model is created based on historical data, eliminating the need for expert assessment of weighting coefficients. Machine learning algorithms transform data into knowledge [Raschka and Mirjalili, 2017], which is available to agents in decision-making.

2 Literature review

In ABMS models representing business systems and processes, agents most often represent people and their behaviour [Gilbert and Troitzsch, 2005]. The main assumption is that agents can credibly model humans and their social interactions at a reasonable level of abstraction [Čavoški, 2016]. For this purpose, agents' essential characteristics, autonomy, and ability to act independently and adapt to given circumstances are used. These features are often cited in the literature as the main justification and advantage for the application of ABMS instead of other simulation methodologies [Macal and North, 2010, Wooldridge and Jennings, 1995].

2.1 Trends in the application of agent-based modelling and simulation

Since its first appearance in the early 1970s, the exploration of the ABMS literature has seen vigorous development, especially in the last decade, owing to the availability of various ABMS software [Marković and Zornić, 2016]. Bibliometric analysis geared towards a review of literature productivity and an observation of the trends in ABMS is presented in this section of the paper. Data is collected from the Web of Science platform [Clarivate Analytics, 2020b] (all sources), using the following search configuration:

TI = (agent-based OR individual-based OR multi-agent OR multiagent OR ABM) AND TI = (model* OR simulat*)*
Indexes = SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI
Timespan = All years

Data is collected for 10,286 scientific papers. The increase in the number of papers is evident, proving that this simulation approach's popularity in the scientific community is rising. Furthermore, analysing the research area in which ABMS is applied, we can provide one more piece of evidence that ABMS is suitable for application in various areas (Table 1, research areas with at least 100 papers in the 2010-2019 period). With almost half of the papers, computer science and engineering are dominant research areas, while environmental sciences & ecology, business & economics, and automation & control systems follow.

Research area	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**Total (2010-2019)**
Computer Science	260	250	290	296	313	368	297	375	330	243	**3022**
Engineering	126	153	180	215	177	220	207	252	201	154	**1885**
Environmental Sciences & Ecology	40	43	54	71	59	62	69	93	87	110	**688**
Business & Economics	31	38	36	30	43	51	61	80	69	61	**500**
Automation & Control Systems	26	22	41	38	51	61	46	47	46	27	**405**
Operations Research & Management Science	48	60	41	38	30	33	38	36	36	23	**383**
Science & Technology - Other Topics	22	10	14	27	38	29	45	58	64	51	**358**
Mathematics	36	25	34	34	35	40	25	41	52	32	**354**
Social Sciences - Other Topics	26	19	17	25	39	40	42	53	40	23	**324**
Transportation	9	16	15	25	33	24	32	39	38	32	**263**
Energy & Fuels	5	10	5	9	29	27	23	27	31	36	**202**
Mathematical & Computational Biology	34	17	12	13	18	16	17	22	20	27	**196**
Physics	15	10	16	17	13	18	20	25	21	19	**174**

Telecommunications	6	5	13	20	6	29	25	20	18	24	**166**
Robotics	3	1	3	6	17	62	17	19	19	7	**154**
Geography	9	14	10	15	11	11	18	9	9	13	**119**
Public, Environmental & Occupational Health	2	7	3	5	6	17	15	12	20	16	**103**

Table 1: Trends in ABMS research area

Among other data, KeyWords Plus [Clarivate Analytics, 2020a] are collected, too. These keywords are words or phrases that often appear in the titles of papers referenced in a particular article. The most common keywords for the period from 2010 to 2019 are given in Table 2. If we disregard the expected keywords, such as dynamics, simulation, systems, behaviour, other keywords describe the scope of application of ABMS.

Keyword	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**Total (2010-2019)**
Dynamics	31	34	26	46	55	58	60	60	67	68	**505**
Simulation	21	30	31	35	43	62	51	70	80	71	**494**
Systems	22	25	18	25	44	44	52	46	65	52	**393**
Behavior	11	20	17	26	31	24	39	27	38	31	**264**
Management	25	19	20	17	22	26	28	31	34	40	**262**
Model	13	13	18	28	22	19	29	29	32	22	**225**
Networks	8	10	11	15	24	16	21	32	33	14	**184**
Framework	5	7	6	10	10	18	13	31	32	38	**170**
Protocol	3	4	14	14	17	23	21	19	23	13	**151**
Design	5	11	10	11	11	11	21	19	25	23	**147**
System	4	2	10	15	16	16	17	13	25	26	**144**
Growth	9	13	10	13	19	14	14	19	13	19	**143**
Impact	5	7	5	7	14	6	11	22	26	28	**131**
Evolution	9	6	8	12	17	18	13	24	12	20	**139**
Performance	2	3	7	7	7	8	15	21	19	22	**111**
Strategies	8	10	6	10	11	9	8	11	14	22	**109**

Optimization	1	4	1	7	9	10	16	26	16	14	**104**
Diffusion	1	7	3	14	8	9	13	18	15	13	**101**

Table 2: ABMS keywords trends

2.2 Utility function in ABMS

Utility function has a long tradition in economics and management sciences. It has been used to capture individuals' preferences, welfare, and the whole economy [Dantzig et al., 1988, Poterba et al., 1986]. This function can combine multiple returns to formalise behaviour and decision-making, such as choosing education while maximising both economic and social returns to education [Jæger, 2007]. When there is a need to weigh some costs and benefits and make a decision, a utility function can be employed, even when choosing a medication [van Dam et al., 2020].

Utility function, as well as global or local utility thresholds, are often used in agent-based simulation models where agents have to make decisions during the interaction. These functions are used to represent the agent's objectives and shape their behaviour.

One of the most significant issues with using the utility function in agent-based simulation models is how this function and the utility threshold are defined. This process usually requires the selection of variables that will be included in a function and then evaluating specific parameters and weights by experts, which is often not easy to do objectively, so it is based on subjective assessment.

The most recent utility function applications in ABMS are presented in Table 3. Most of the models use the maximum utility as a decision criterion, while some employ utility maximisation and utility threshold. All the authors are giving their best to define utility functions using data-driven, objective values. Unfortunately, in order to capture agents' stochastic behaviour using utility function, some sensitivity parameters and weight coefficients have to be used. All of the models listed in Table 3 have at least one of those estimated.

Paper	Model topic	Utility function components	Decision criteria	Estimated function part
[Hu et al., 2020]	Goods exchange	Wealth, goods price, satisfaction	Maximum utility	Preference parameters
[Tomasiello et al., 2020]	Job accessibility and its inequalities	Job accessibility; Neighbourhood status	Utility maximisation	Weight parameter
[Zhang and Zheng, 2019]	Consumer purchase behaviour	Commodity quality, price, and promotion	Maximum utility	Weights and sensibility parameters
[Badraoui et al., 2017]	Impact of qualitative	Risk aversion; Individual	Utility maximisation	Sensitivity parameters

	characteristics on market equilibrium	wealth; Quality, asset, nation, and individual compliance		
[Roozmand et al., 2011]	Consumer decision making based on power distance and personality	Product social status; Product social responsibility; Product price	Maximum utility	Weights
[Čavoški and Marković, 2016]	Consumer behaviour on B2C e-commerce sites	Product price; Marketing campaign effects; Demographic attributes; Website rating	Maximum utility and utility threshold	Weights and sensibility parameters
[Zhang and Zhang, 2007]	Consumer purchase decision-making and the decoy effect	Brand quality; Brand price; Advertising; Brand follower tendency; Exerted brand influence	Maximum utility	Sensitivity, susceptibility
[Li-xia et al., 2009]	Automated negotiation system	Proposed price	Maximum utility	Weights

Table 3: A survey of utility function usage in ABMS

Čavoški [Čavoški, 2016, Čavoški and Marković, 2016] uses a utility function to model the decision-making process of online purchasing. In combination with the utility function, the utility threshold is also used – if the value of the utility function is below the defined threshold, the purchase decision will be unfavourable. The utility function encompasses a wide range of variables, and each of them requires determining weight coefficients. Similarly, Roozmand et al. [Roozmand et al., 2011] used a utility function to model consumer decision-making processes, including consumers' culture, personality, and human needs. Like most of the other ones, both of these papers require choosing variables comprising utility function and weight coefficients or sensitivity for specific variables. Zhang and Zheng [Zhang and Zheng, 2019] used a similar approach while investigating the influence of quality, price, and promotion on consumers' purchase behaviour.

Hu et al. [Hu et al., 2020] modelled price evolutions of goods exchange in a multi-agent market, using the kinetic theory. They employed multiple utility functions involving agents' wealth, goods price, and satisfaction to describe agents' decision-making in the process of goods exchange.

Tomasiello et al. [Tomasiello et al., 2020] presented an agent-based model for exploring job accessibility inequalities among different social groups. They focused their investigation on the impact of public transport and land-use policies on the residential location of the working population and their accessibility to job opportunities. Their utility function includes job accessibility and neighbourhood status, with weight parameter estimated.

Badraoui et al. [Badraoui et al., 2017] developed a financial utility function that involves compliance to a qualitative, ethical aspect. Their goal is to prove that individual utility may depend on parameters other than wealth and risk aversion. After defining the function, they conducted simulation experiments in NetLogo and showed the function's effectiveness. The framework proposed in our paper makes using qualitative variables in the decision-making process even easier by using one of many machine learning algorithms supporting qualitative variables.

Analysing papers that employ utility function in ABMS, one can identify two main issues: (1) selecting variables that will make up utility functions; (2) determining weight and sensitivity coefficients and calibrating the function.

A step towards data-driven replacement of utility function should be made in order to overcome the aforementioned shortcomings of the traditional utility function. Not long ago, Heppenstall et al. [Heppenstall et al., 2021] wrote about future developments of agent-based models using machine learning and said that "... a more fruitful path may lie in facilitating ways for agents to learn and make decisions for themselves" [Heppenstall et al., 2021]. Exactly this approach is employed in the methodological framework proposed in this paper – the traditional utility function is replaced with a machine learning algorithm, thus creating a hybrid model.

2.3 Machine learning in agent-based simulation models

People generate more and more data every day [Desjardins, 2019], which lead to the rapid development of statistical learning [James et al., 2013]. It is estimated that by the end of 2020, the digital universe consisted of 44 zettabytes (1021 bytes) of data, and by the end of 2025, 463 exabytes (1018 bytes) of data will be generated daily [Desjardins, 2019]. The growth in data availability has also led to the development of the ways in which they are used. Data can be used to come to the description of certain phenomena, but also to certain conclusions and decision-making models, by applying more advanced methods of data analysis, such as statistical or machine learning. Companies use the digital footprint of individuals to tailor a marketing campaign to an individual with the help of psychological targeting [Matz et al., 2017]. Machine learning can also be used for: the prevention of certain diseases [Barrett et al., 2013]; risk assessment and diagnosis in patients [Canullo et al., 2016]; reducing congestion and increasing the safety of traffic participants [Shi and Abdel-Aty, 2015]; demand forecasting [Song and Liu, 2017]; credit risk assessment [Pérez-Martín et al., 2018]; identifying clients who will leave the company [Hung et al., 2006] and more.

A review of the relevant literature found that other authors also used machine learning algorithms when building agent-based simulation models. The difference is in the part of the model that is modelled using a machine learning algorithm. In this paper, a machine learning algorithm replaces the utility function, while other authors introduce different possible applications.

Lamperti et al. [Lamperti et al., 2018] use machine learning and iterative sampling to calibrate ABM with real data. The results obtained show that machine learning enables the precise representation of the system by the model and drastically reduces the time required for searching large parameter space and calibration. A similar approach is used by Kavak et al. [Kavak et al., 2018]. Namely, they use machine learning and available data to generate agents and assign values to appropriate attributes.

On the other hand, it is possible to apply machine learning to detect the influence of input variables on the outputs of the model [R. Vahdati et al., 2019]. While other authors [Hayashi et al., 2016] compare the results obtained from ABMS with the results obtained by the machine learning model and, on that basis, perform additional calibration of ABM.

3 A methodological framework for the integration of machine learning algorithms into agent-based simulation models

A framework for the hybrid agent-based model that aims to overcome some of the elaborated problems related to the usage of a utility function is developed. In the proposed approach, a machine learning algorithm is incorporated in the place of a utility function, thus providing a possibility to use various algorithms.

Creating a hybrid model begins with building an agent-based simulation model (Figure 1). The model is built considering aspects suggested by Macal and North [Macal and North, 2010]:

- Specific problem to be solved by the ABMS;
- Definition of agents and their static/dynamic attributes;
- Design of an environment and the way agents interact with it;
- Agents' behaviours;
- Agents' mutual interactions;
- Availability of data;
- Model validation, especially the agent behaviours.

In the next step, a selected machine learning model is built based on historical data. The machine learning model is chosen based on available data and the needed type of prediction.

Those two models, agent-based simulation and machine learning model are further used in synergy. As agents interact, the values of their attributes change. When the time for agents to decide comes, instead of decision-making based on the utility function, attribute values are passed through the machine learning model. The predicted value represents the agent's decision.

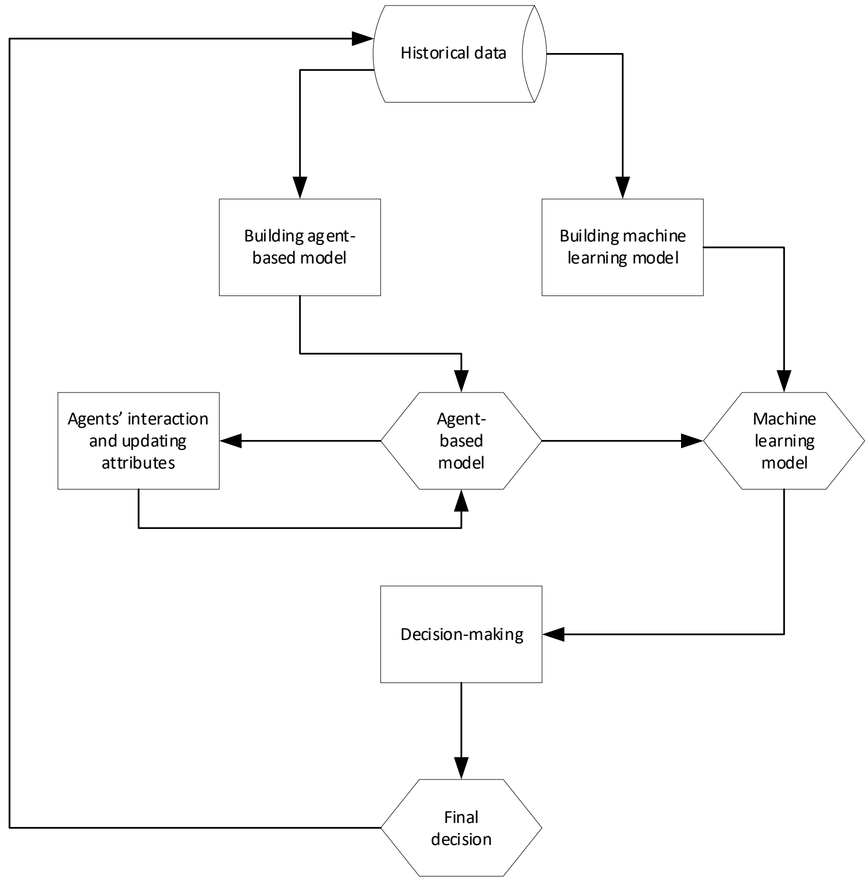

Figure 1: Methodological framework for superseding utility function with a machine learning model

The presented hybrid approach overcomes some of the identified problems related to the usage of the utility function. First of all, machine learning algorithms have built-in mechanisms for determining the importance of variables, so the expert does not have to decide which variables will be observed in the decision-making process or determine the values of weight coefficients. Also, machine learning algorithms, such as the decision tree, have a built-in ability to use qualitative variables and decide based on the value of only one variable, i.e., a single split can sometimes be sufficient for making a particular decision.

The proposed methodological framework introduces improvements both for ABMS and machine learning. Decision-making in ABMS is impartial and objective, and at the same time, machine learning algorithms are enabled with scenario analysis based on agents' interactions. Using a machine learning model in place of the utility function is more convenient when historical data is available. In some cases, machine learning can be used as a tool for parametrising the utility function. On the other side, when historical data is not available, modeller has to rely on using utility functions for capturing agents' objectives.

4 Application of the proposed methodological framework: A case of churn in a telecommunications company

One of the common machine learning applications is churn prediction, often used in the telecommunications industry [Ahmad et al., 2019, Pamina et al., 2020, Qureshi et al., 2013, Ullah et al., 2019]. Customer churn is one of the most challenging problems that affect revenue and customer base in mobile telecom operators [Alboukaey et al., 2020]. Publicly available data [IBM, 2018] and some common sense behaviour assumptions were used to build the hybrid model.

The available database contains 20 variables (18 out of 20 variables are used) that describe 7043 clients. Eleven clients had missing data for some of the variables, and they were excluded from the analysis. The variables used are given in Table 4.

Variable	**Description**
gender	Client gender (male, female)
SeniorCitizen	Is the client retired (yes, no)
Partner	Does the client have a partner (yes, no)
Dependents	Does the client have dependents (yes, no)
tenure	The number of months a client is in the company
PhoneService	Does the client have a landline (yes, no)
MultipleLines	Does the client have multiple landlines (yes, no, no phone)
InternetService	The type of internet that the client uses (DSL, optical, no internet)
OnlineSecurity	Does the client have online protection (yes, no, no internet)
OnlineBackup	Does the client have an online data backup (yes, no, no internet)
DeviceProtection	Does the client have device insurance (yes, no, no internet)
TechSupport	Does the client have technical support (yes, no, no internet)
StreamingTV	Does the client have streaming TV (yes, no, no internet)
StreamingMovies	Does the client have streaming movies (yes, no, no internet)
Contract	Contractual obligation of the client (month-to-month, one year, two years)
PaperlessBilling	Does the client use an electronic invoice (yes, no)
PaymentMethod	Method of payment used by the client (electronic check, mailed check, bank transfer (automatic), credit card (automatic))
Churn	Whether the client leaves the company or not (yes, no)

Table 4: Variables used in the modelling process

Of the 7032 observed clients, 27%, i.e. 1869, plan to terminate the company's services. These clients are in the focus of the proposed hybrid approach. Building a hybrid model begins with creating an agent-based simulation model. In this process, historical data, known elements of the system, connections between them and certain assumptions were used.

The agent-based model includes three types of agents:

1. Agents representing clients – *client agents*;
2. Agents representing offers for clients from the telecommunications company sales department – *offer agents*;
3. Agents that represent connections between clients – *connection agents*.

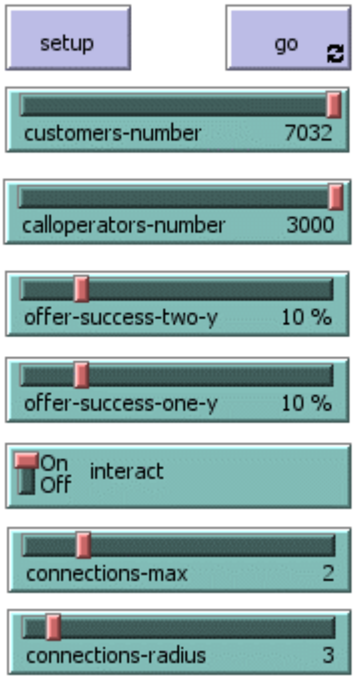

Figure 2: The user interface for controlling the simulation experiment

Client agents' attributes values are assigned from an available data set. After all the *client agents* are created, a certain number of connections are established between them at random (it is possible to adjust the number of connections via the graphical interface – Figure 2). These connections represent the connections of clients with family and close friends, with whom they exchange experiences on telecommunication services. Based on the experience exchange, clients will potentially change their behaviour – agents whose contract expires exchange opinions with their connections. If the majority (more than 50%) has a specific type of contract, then the observed agent chooses the same type of contract. This influence on the agents' behaviour is visible in a simulation experiment in which interactions between clients are observed.

The next step is creating a machine learning model based on the previously mentioned historical data. The first model is built using random forest algorithm (Model 1). Additionally, the proposed approach is applied using the decision tree machine learning model (Model 2) in order to present the approach's ability to employ different machine learning algorithms. In the end, the comparison to the model with utility function is made (Model 3).

Python [Python, 2020] in the *JupyterLab* [Jupyter, 2020] working environment was used to implement a random forest and decision tree, and after adequate preparation of the algorithm, the code was translated into *NetLogo* [Wilensky, 1999]. Python packages *scikit-learn* [Pedregosa et al., 2011], *pandas* [Pandas Development Team, 2020] and *numpy* [van der Walt et al., 2011] were used in data analysis.

The data set is divided into a training and a test set with 75% and 25% of the cases, respectively. The random forest algorithm from scikit-learn package was initialised with 1000 trees, and an accuracy of 77.35% on the test set was obtained, while the decision tree algorithm from the same package obtained an accuracy of 69.97% on the test set.

Variable	**Average importance**
tenure	27.26%
Contract	11.26%
PaymentMethod	8.50%
OnlineSecurity	6.09%
TechSupport	5.84%
InternetService	5.83%
MultipleLines	4.00%
OnlineBackup	3.91%
gender	3.71%
DeviceProtection	3.49%
StreamingMovies	3.48%
StreamingTV	3.46%
PaperlessBilling	3.40%
Partner	3.19%
SeniorCitizen	2.94%
Dependents	2.75%
PhoneService	0.90%

Table 5: Average variables' Gini importance

After learning the random forest model, the average importance of the variables for all 1000 trees was calculated. The importance of the variable is computed as the normalised total reduction of the criterion brought by the variable – also known as the Gini importance [scikit-learn, 2022]. The highest level of importance, just over 27%, has a variable *time that the client spent in the company* (tenure). It is followed by the variable *contract* (a current type of client's contract), with 11%, followed by the variable *payment method* and other variables (Table 5).

It is impossible to influence how long someone will be a client directly, so it is interesting to see how other variables in the model will affect this period. Here, the *contract* stood out as the next most important variable, so it is presented in more detail for those clients who plan to leave the company (Figure 3). Interestingly, over 88% of clients who plan to leave the company have a month-to-month contract. In further model building and simulation experiments, the behaviour of these clients will be in our focus.

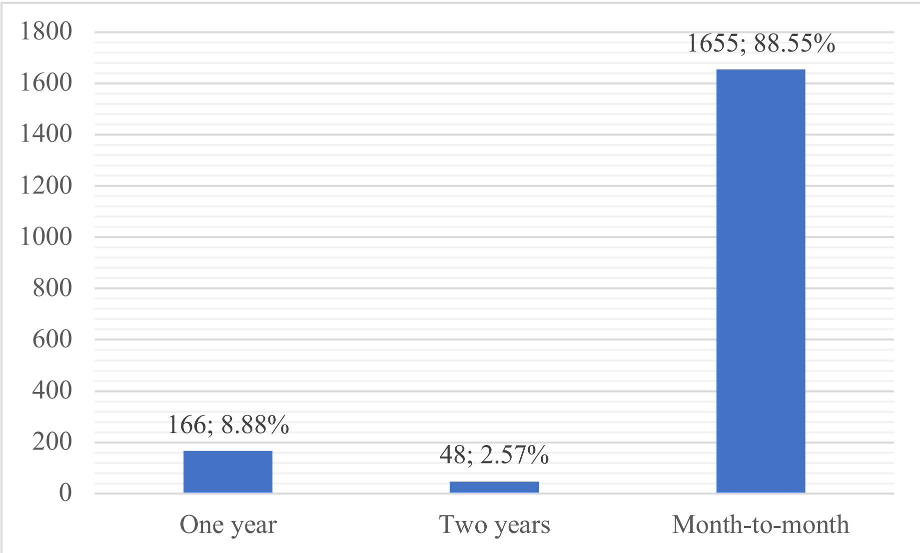

Figure 3: Type of contract among clients that are leaving the company

In addition to the previously described *client agents* and the links between them, the model also includes *offer agents* – representing personalised offers for clients from the telecommunications company's sales department. *Offer agents* are tasked with targeting *client agents*, whom the machine learning model depicted as those who will potentially leave the company. The sales department is offering the *client agent*s a one-year or two-year contract. Sales department performance for each contract type can be adjusted via a graphical interface (Figure 2). One of the assumptions of the model is that there is a sufficient number of *offer agents* to provide the offer to all *client agents* simultaneously. This is justified by the real possibility of delivering a special offer to many clients in a short time using various communication channels, such as e-mail, text messages, messages on social networks, TV commercials, and others.

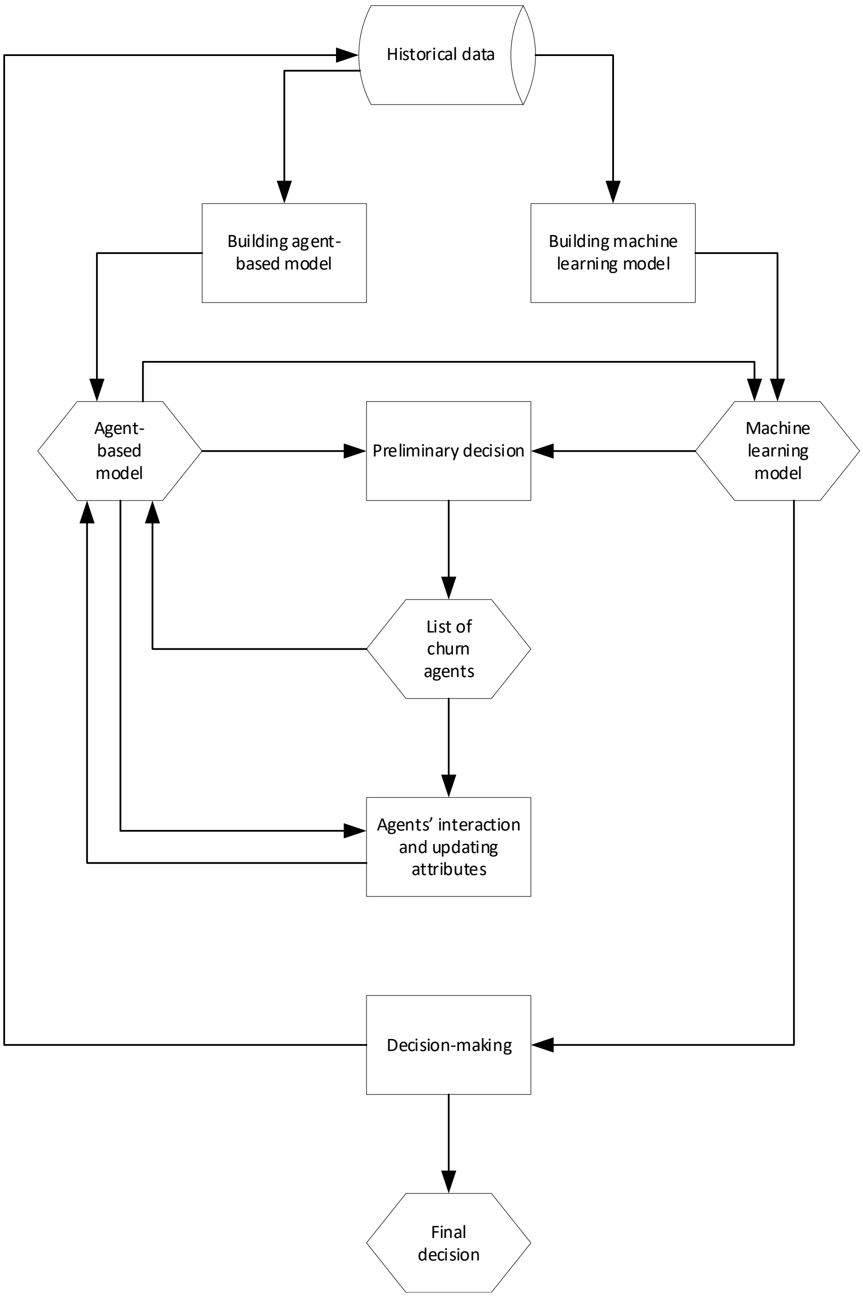

Figure 4: Agent-based model roadmap for a case study of churn in a telecommunications company (adapted from Figure 1)

Figure 4 shows agents' decision-making process, adjusted based on the proposed methodological framework for superseding the utility function with machine learning algorithms (Figure 1). Two developed models, an agent-based simulation model and a machine learning model, are further used in synergy. The agents' attributes are passed through the machine learning model resulting in a preliminary decision for each agent. Based on the previous procedure, a list of potential candidates to decide to leave the company is obtained.

This list can direct specific actions only on those agents who would make an undesirable decision for the company (churn). The next step is to initiate agents' interaction. During the interaction, agents' attributes are updated and run through the machine learning model, thus representing the agents' decision-making process.

In the presented simulation experiments, changes in clients' behaviour were observed by influencing one variable – type of contract. Of course, the model is easy to expand by observing the broader set of variables.

4.1 Reference model

In addition to two models that employ machine learning for agents' decision-making, the utility-function-based model is developed to be used as a reference (Model 3).

The utility function is defined using variables' Gini importance (Table 5) as a starting point. The six most important variables are selected (*tenure, Contract, PaymentMethod, OnlineSecurity, TechSupport, InternetService*), and the logistic regression model is built using scikit-learn package with an accuracy of 77.19% on the test set. Logistic regression model coefficients are used with dummy variables (all except *tenure*) to compute l:

$$
\begin{aligned}
l = {} & 0.00072372 \\
& -0.06007908 \cdot tenure \\
& -0.07378934 \cdot InternetService\ [DSL] \\
& +0.19842073 \cdot InternetService\ [Fiber\ optic] \\
& -0.12390729 \cdot InternetService\ [No] \\
& +0.15638822 \cdot OnlineSecurity\ [No] \\
& -0.12390729 \cdot OnlineSecurity\ [No\ internet\ service] \\
& -0.03175683 \cdot OnlineSecurity\ [Yes] \\
& +0.15922284 \cdot TechSupport\ [No] \\
& -0.12390729 \cdot TechSupport\ [No\ internet\ service] \\
& -0.03459145 \cdot TechSupport\ [Yes] \\
& +0.10712535 \cdot Contract\ [Month-to-month] \\
& -0.0449842 \cdot Contract\ [One\ year] \\
& -0.06141706 \cdot Contract\ [Two\ year] \\
& -0.0108933 \cdot PaymentMethod\ [Bank\ transfer\ (automatic)] \\
& -0.02730723 \cdot PaymentMethod\ [Credit\ card\ (automatic)] \\
& +0.13519598 \cdot PaymentMethod\ [Electronic\ check] \\
& -0.09627135 \cdot PaymentMethod\ [Mailed\ check]
\end{aligned}
$$

l is then used in the following formula in order to get the probability [Kleinbaum and Klein, 2010]:

$$UF = \frac{1}{1 + e^{-l}}$$

Based on probability, clients are classified as those who will leave the company (p>0.5) or not (p≤0.5).

4.2 Experiment 1: No interaction between client agents

In this simulation experiment, all *client* and *offer agents* were observed. *Offer agents* had a success rate of 10% for both one-year and two-year contracts (Figure 5). Solely this type of interaction – between client and offer agents was analysed. Results of this experiment indicated that the offer success rate was the only influential factor.

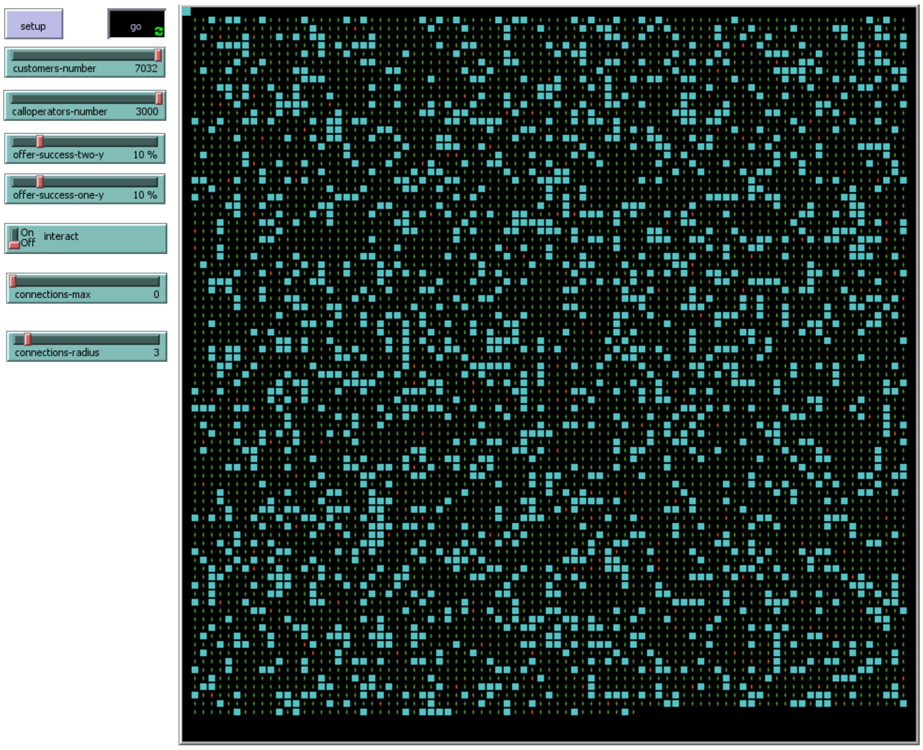

Figure 5: NetLogo world – Experiment 1: no interaction between client agents

Model 1: The random forest machine learning model estimated that 1771 clients would leave the company, and 1615 had a one-month contract. These clients were given the special offer, and 434 accepted a one-year/two-year contract, after which 428 of them were classified as risk-free, i.e., it was estimated that they would not leave the company.

Model 2: The decision tree machine learning model estimated that 1858 clients would leave the company, and 1635 had a one-month contract. These clients were given the special offer, and 624 accepted a one-year/two-year contract, after which 547 of them were classified as risk-free, i.e., it was estimated that they would not leave the company.

Model 3: The machine learning parametrised utility function model estimated that 1234 clients would leave the company, and 1230 had a one-month contract. These clients were given the special offer, and 240 accepted a one-year/two-year contract, after which 174 of them were classified as risk-free, i.e., it was estimated that they would not leave the company.

4.3 Experiment 2: Interaction between client agents

This simulation experiment is the more complex one, as the client interactions are now observed in the model. The number of connections that one client agent makes with surrounding client agents is two (Figure 6).

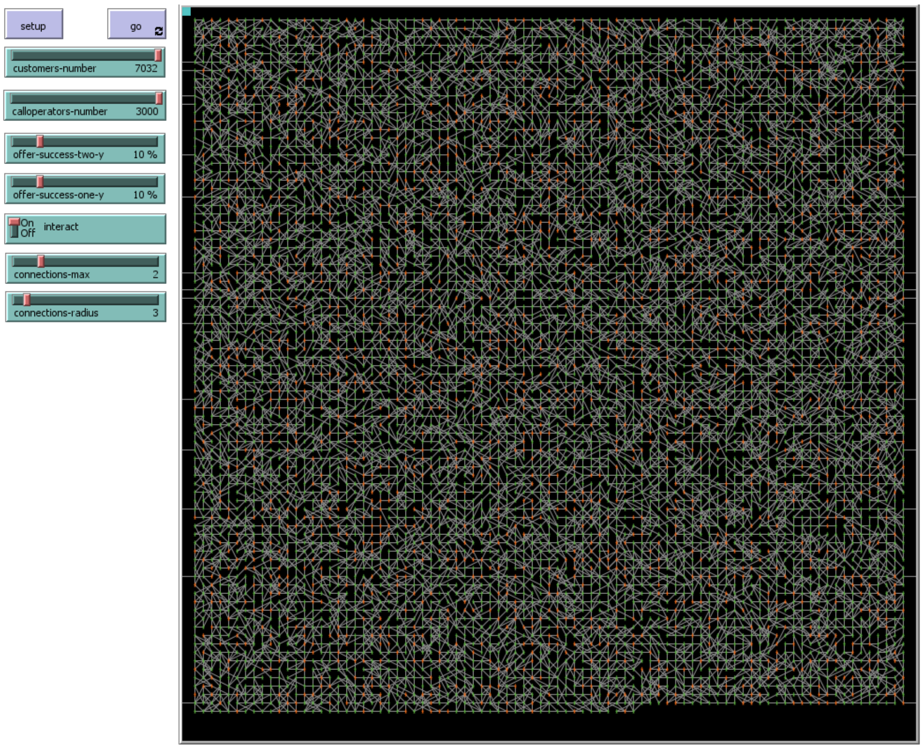

Figure 6: NetLogo world – Experiment 2: interaction between client agents

Based on this simulation experiment results, we conclude that the most significant impact on the number of clients switching from a monthly to a one-year/two-year contract has the offer from the telecommunications company's sales department (Figure 6). Besides, there is a change in client behaviour based on interaction with related clients.

Model 1: After contacting 1615 clients identified as those planning to leave the company in the first month, another 53 clients with a one-month contract were identified and contacted in the next 36 months. A total of 587 clients have been classified as risk-free.

Model 2: After contacting 1642 clients identified as those planning to leave the company in the first month, another 85 clients with a one-month contract were identified and contacted in the next 36 months. A total of 688 clients have been classified as risk-free.

Model 3: After contacting 1230 clients identified as those planning to leave the company in the first month, only one additional client with a one-month contract was identified and contacted in the next 36 months. A total of 184 clients have been classified as risk-free.

4.4 Discussion

Analysing different approaches in modelling agents' decision-making, one will identify the pros and cons of each approach.

The utility function can be employed even when there is no historical data, which is especially useful for scientific purposes and when an expert's opinion plays an important role. On the other side, formulating an objective utility function is not easy to accomplish. Selecting important variables, defining variables' importance, utility threshold – those are all potential elements comprising utility function and its deployment in the model.

Using a machine learning model to model an agent's decision-making is much more convenient when historical data related to a decision is available. Various machine learning algorithms are available nowadays, allowing a modeller to apply them to all kinds of data. Those algorithms are providing agents with an objective, data-driven tool for decision-making, but only if historical data is available.

Models 1 and 2 observed in this section of the paper employ machine learning algorithms, while Model 3 employs utility function. Observing their performance in the ABM setting, Model 1 had 93.03% accuracy, Model 2 91.20% and Model 3 78.80%. We can see that in addition to being more convenient to model, machine-learning-based decision-making also yielded higher accuracy, thus better-representing agents' behavioural patterns.

5 Conclusion

The complexity of the economic [Cárdenas et al., 2018] and social, technical and other systems grow with their development. As a consequence, there was a need for analysis and a better understanding of how they work. The application of modelling and simulations can give a good insight into the behaviour of such systems. The development of the observed system model enables the analysis of how the system is functioning and increase in its understanding. Agent-based modelling is used to model complex systems in which the interaction of their key elements is present.

The utility function is often used in agent-based simulation models in which agents are required to make decisions. This function is calculated based on the values of the agent's attributes and is most often used in two ways: (1) the agent maximises its value; (2) the agent decides on the outcome for which the utility function exceeds the defined threshold. The choice of variables that will be included in the utility function, the weighting coefficients, and the threshold of the utility function often require an expert-driven subjective assessment.

As data is becoming increasingly available to us, it should be exploited to a greater extent in the model building process. Namely, algorithms that can detect patterns in data, learn using historical data and incorporate obtained knowledge into the decision-making process are machine learning algorithms.

This paper presents a hybrid approach for successfully integrating machine learning algorithms into agent-based simulation models. Available historical data is used to build both models, and then these models are used in synergy. Agents interact, their attributes change, and then a machine learning model is used for decision-making. The hybrid approach eliminates the need to define the utility function and all its elements.

However, some limitations should be noted. Historical datasets are not always available for the observed system; thus, machine learning algorithms cannot be utilised. Additionally, expert opinion in some situations can be essential and should be used together with machine learning algorithms for modelling agents' decision-making.

References

[Abbott, 2006] Abbott, R.: 'Emergence explained: Abstractions: Getting epiphenomena to do real work'; Complexity, 12, 1 (2006), 13–26. https://doi.org/10.1002/cplx.20146

[Ahmad et al., 2019] Ahmad, A. K., Jafar, A., Aljoumaa, K.: 'Customer churn prediction in telecom using machine learning in big data platform'; Journal of Big Data, 6, 1 (2019), 28. https://doi.org/10.1186/s40537-019-0191-6

[Alboukaey et al., 2020] Alboukaey, N., Joukhadar, A., Ghneim, N.: 'Dynamic behavior based churn prediction in mobile telecom'; Expert Systems with Applications, 162 (2020), 113779. https://doi.org/10.1016/j.eswa.2020.113779

[Ali et al., 2019] Ali, A., Bajwa, I., Kazmi, R.: 'High-performance simulation of drug release model using finite element method with CPU/GPU platform'; JUCS-Journal of Universal Computer Science, 25 (2019), 1261.

[Badraoui et al., 2017] Badraoui, S., Bensaid, K., El Haloui, A. M., Aboulaich, R.: 'Impact of Qualitative Characteristics on Market Equilibrium: Simulation with Agent-Based Modeling'; Review of Pacific Basin Financial Markets and Policies, 20, 02 (2017), 1750013. https://doi.org/10.1142/S0219091517500138

[Barrett et al., 2013] Barrett, M. A., Humblet, O., Hiatt, R. A., Adler, N. E.: 'Big Data and Disease Prevention: From Quantified Self to Quantified Communities'; Big Data, 1, 3 (2013), 168–175. https://doi.org/10.1089/big.2013.0027

[Canullo et al., 2016] Canullo, L., Tallarico, M., Radovanovic, S., Delibasic, B., Covani, U., Rakic, M.: 'Distinguishing predictive profiles for patient-based risk assessment and diagnostics of plaque induced, surgically and prosthetically triggered peri-implantitis'; Clinical Oral Implants Research, 27, 10 (2016), 1243–1250. https://doi.org/10.1111/clr.12738

[Cárdenas et al., 2018] Cárdenas, J. P., Vidal, G., Urbina, C., Olivares, G., Rodrigo, P., Fuentes, M.: 'Social Crises: Signatures of Complexity in a Fast-Growing Economy'; Complexity, 2018 (2018), 1–11. https://doi.org/10.1155/2018/9343451

[Čavoški, 2016] Čavoški, S.: 'Simulacioni modeli zasnovani na agentima kao podrška odlučivanju u elektronskom poslovanju'; Fakultet organizacionih nauka, Univerzitet u Beogradu (2016).

[Čavoški and Marković, 2016] Čavoški, S., Marković, A.: 'Agent-based modelling and simulation in the analysis of customer behaviour on B2C e-commerce sites'; Journal of Simulation (2016), 1–11. https://doi.org/10.1057/s41273-016-0034-9

[Clarivate Analytics, 2020a] Clarivate Analytics: 'KeyWords Plus generation, creation, and changes'; (2020a). Retrieved 31 August 2020, from https://support.clarivate.com/ScientificandAcademicResearch/s/article/KeyWords-Plus-generation-creation-and-changes?language=en_US

[Clarivate Analytics, 2020b] Clarivate Analytics: 'Web of Science'; (2020b). Retrieved 30 August 2020, from https://apps.webofknowledge.com

[Dantzig et al., 1988] Dantzig, G. B., McAllister, P. H., Stone, J. C.: 'Deriving a Utility Function For the U.S. Economy'; (1988, April 1). Retrieved from https://apps.dtic.mil/sti/citations/ADA197135

[Desjardins, 2019] Desjardins, J.: 'How much data is generated each day? | World Economic Forum'; (2019). Retrieved 14 September 2020, from https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

[Forrester, 1961] Forrester, J.: 'Industrial Dynamics'; Cambridge, MA: The M.I.T. Press (1961).

[Geng et al., 2020] Geng, B., Brahma, S., Wimalajeewa, T., Varshney, P. K., Rangaswamy, M.: 'Prospect Theoretic Utility Based Human Decision Making in Multi-Agent Systems'; IEEE Transactions on Signal Processing, 68 (2020), 1091–1104. https://doi.org/10.1109/TSP.2020.2970339

[Gilbert and Troitzsch, 2005] Gilbert, G. N., Troitzsch, K. G.: 'Simulation for the social scientist'; Open University Press (2005).

[Goldsman et al., 2010] Goldsman, D., Nance, R. E., Wilson, J. R.: 'A brief history of simulation revisited'; In Proceedings - Winter Simulation Conference (2010). https://doi.org/10.1109/WSC.2010.5679129

[Hayashi et al., 2016] Hayashi, S., Prasasti, N., Kanamori, K., Ohwada, H.: 'Improving Behavior Prediction Accuracy by Using Machine Learning for Agent-Based Simulation'; (2016), 280–289. https://doi.org/10.1007/978-3-662-49381-6_27

[Heppenstall et al., 2021] Heppenstall, A., Crooks, A., Malleson, N., Manley, E., Ge, J., Batty, M.: 'Future Developments in Geographical Agent-Based Models: Challenges and Opportunities'; Geographical Analysis, 53, 1 (2021), 76–91. https://doi.org/10.1111/gean.12267

[Hu et al., 2020] Hu, C., Lai, S., Lai, C.: 'Investigations to the price evolutions of goods exchange with CES utility functions'; Physica A: Statistical Mechanics and Its Applications, 549 (2020), 123938. https://doi.org/10.1016/j.physa.2019.123938

[Hung et al., 2006] Hung, S.-Y., Yen, D. C., Wang, H.-Y.: 'Applying data mining to telecom churn management'; Expert Systems with Applications, 31, 3 (2006), 515–524. https://doi.org/10.1016/j.eswa.2005.09.080

[IBM, 2018] IBM: 'Telco Customer Churn'; (2018). Retrieved from https://www.kaggle.com/blastchar/telco-customer-churn/version/1

[Jæger, 2007] Jæger, M. M.: 'Economic and Social Returns To Educational Choices'; Rationality and Society, 19, 4 (2007), 451–483. https://doi.org/10.1177/1043463107083739

[Jahangirian et al., 2010] Jahangirian, M., Eldabi, T., Naseer, A., Stergioulas, L. K., Young, T.: 'Simulation in manufacturing and business: A review'; European Journal of Operational Research, 203, 1 (2010), 1–13. https://doi.org/10.1016/j.ejor.2009.06.004

[James et al., 2013] James, G., Witten, D., Hastie, T., Tibshirani, R.: 'An Introduction to Statistical Learning'; (8th ed., Vol. 103). New York, NY: Springer New York (2013). https://doi.org/10.1007/978-1-4614-7138-7

[Jupyter, 2020] Jupyter: 'Project Jupyter | Home'; (2020). Retrieved 29 September 2020, from https://jupyter.org/

[Kavak et al., 2018] Kavak, H., Padilla, J. J., Lynch, C. J., Diallo, S. Y.: 'Big data, agents, and machine learning: Towards a data-driven agent-based modeling approach'; In Simulation Series (2018).

[Kleinbaum and Klein, 2010] Kleinbaum, D. G., Klein, M.: 'Introduction to Logistic Regression'; In Logistic Regression. Statistics for Biology and Health. Springer, New York, NY (2010), 1–39. https://doi.org/10.1007/978-1-4419-1742-3_1

[Lamperti et al., 2018] Lamperti, F., Roventini, A., Sani, A.: 'Agent-based model calibration using machine learning surrogates'; Journal of Economic Dynamics and Control, 90 (2018), 366–389. https://doi.org/10.1016/j.jedc.2018.03.011

[Li-xia et al., 2009] Li-xia, L., Lan, S., Hong, W.: 'Simulation Study of Multi-Agent Based Automated Negotiation System in E-Commerce'; In 2009 Third International Symposium on Intelligent Information Technology Application. IEEE (2009), 419–422. https://doi.org/10.1109/IITA.2009.158

[Macal and North, 2010] Macal, C., North, M.: 'Tutorial on agent-based modelling and simulation'; Journal of Simulation, 4, 3 (2010), 151–162. https://doi.org/10.1057/jos.2010.3

[Marković and Zornić, 2016] Marković, A., Zornić, N.: 'Trends in the Application of Agent-Based Modeling and Simulation'; In Central European Conference on Information and Intelligent Systems. Varaždin, Croatia: University of Zagreb, Faculty of Organization and Informatics Varaždin (2016), 65–70.

[Matz et al., 2017] Matz, S. C., Kosinski, M., Nave, G., Stillwell, D. J.: 'Psychological targeting as an effective approach to digital mass persuasion'; Proceedings of the National Academy of Sciences, 114, 48 (2017), 12714–12719. https://doi.org/10.1073/pnas.1710966114

[Mourtzis et al., 2014] Mourtzis, D., Doukas, M., Bernidaki, D.: 'Simulation in Manufacturing: Review and Challenges'; Procedia CIRP, 25 (2014), 213–229. https://doi.org/10.1016/j.procir.2014.10.032

[Pamina et al., 2020] Pamina, J., Beschi Raja, J., Sam Peter, S., Soundarya, S., Sathya Bama, S., Sruthi, M. S.: 'Inferring Machine Learning Based Parameter Estimation for Telecom Churn Prediction'; (2020), 257–267. https://doi.org/10.1007/978-3-030-37218-7_30

[Pandas Development Team, 2020] Pandas Development Team: 'Pandas 1.0.3'; (2020). https://doi.org/10.5281/zenodo.3715232

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al.: 'Scikit-learn: Machine learning in Python'; Journal of Machine Learning Research (2011).

[Pérez-Martín et al., 2018] Pérez-Martín, A., Pérez-Torregrosa, A., Vaca, M.: 'Big Data techniques to measure credit banking risk in home equity loans'; Journal of Business Research, 89 (2018), 448–454. https://doi.org/10.1016/j.jbusres.2018.02.008

[Poterba et al., 1986] Poterba, J. M., Rotemberg, J. J., Barnett, W., Blanchard, O., Fischer, S., Hansen, L.: 'Money in the Utility Function: An Empirical Implementation'; New Approaches to Monetary Economics: Proceedings of the Second International Symposium in Economic Theory and Econometrics (1986), 219–240. https://doi.org/10.3386/W1796

[Prokopenko et al., 2009] Prokopenko, M., Boschetti, F., Ryan, A. J.: 'An information-theoretic primer on complexity, self-organisation, and emergence'; Complexity, 15, 1 (2009), 11–28. https://doi.org/10.1002/cplx.20249

[Python, 2020] Python: 'Welcome to Python.org'; (2020). Retrieved 29 September 2020, from https://www.python.org/

[Qureshi et al., 2013] Qureshi, S. A., Rehman, A. S., Qamar, A. M., Kamal, A., Rehman, A.: 'Telecommunication subscribers' churn prediction model using machine learning'; In Eighth International Conference on Digital Information Management (ICDIM 2013). IEEE (2013), 131–136. https://doi.org/10.1109/ICDIM.2013.6693977

[R. Vahdati et al., 2019] R. Vahdati, A., Weissmann, J. D., Timmermann, A., Ponce de León, M. S., Zollikofer, C. P. E.: 'Drivers of Late Pleistocene human survival and dispersal: an agent-based modeling and machine learning approach'; Quaternary Science Reviews, 221 (2019), 105867. https://doi.org/10.1016/j.quascirev.2019.105867

[Raschka and Mirjalili, 2017] Raschka, S., Mirjalili, V.: 'Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow'; (Second Edi). Birmingham: Packt Publishing (2017).

[Roozmand et al., 2011] Roozmand, O., Ghasem-Aghaee, N., Hofstede, G. J., Nematbakhsh, M. A., Baraani, A., Verwaart, T.: 'Agent-based modeling of consumer decision making process based on power distance and personality'; Knowledge-Based Systems, 24, 7 (2011), 1075–1095. https://doi.org/10.1016/j.knosys.2011.05.001

[scikit-learn, 2022] scikit-learn: 'sklearn.ensemble.RandomForestClassifier — scikit-learn 1.0.2 documentation'; (2022). Retrieved 23 January 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.feature_importances_

[Shi and Abdel-Aty, 2015] Shi, Q., Abdel-Aty, M.: 'Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways'; Transportation Research Part C: Emerging Technologies, 58 (2015), 380–394. https://doi.org/10.1016/j.trc.2015.02.022

[Song and Liu, 2017] Song, H., Liu, H.: 'Predicting Tourist Demand Using Big Data'; In Z. Xiang & D. Fesenmaier (Eds.), Analytics in Smart Tourism Design (2017), 13–29. https://doi.org/10.1007/978-3-319-44263-1_2

[Taylor, 2014] Taylor, S. J.: 'Introducing agent-based modeling and simulation'; In S. J. Taylor (Ed.), Agent-based Modeling and Simulation. Palgrave Macmillan (2014). https://doi.org/10.1057/9781137453648.0003

[Tocher and Owen, 1960] Tocher, K., Owen, D.: 'The automatic programming of simulations'; In J. Banbury & J. Maitland (Eds.), Proceedings of the Second International Conference on Operational Research. English Universities Press: UK (1960), 58–60.

[Tomasiello et al., 2020] Tomasiello, D. B., Giannotti, M., Feitosa, F. F.: 'ACCESS: An agent-based model to explore job accessibility inequalities'; Computers, Environment and Urban Systems, 81 (2020), 101462. https://doi.org/10.1016/j.compenvurbsys.2020.101462

[Ullah et al., 2019] Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., Kim, S. W.: 'A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector'; IEEE Access, 7 (2019), 60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999 [van Dam et al., 2020] van Dam, C. J., Algera, M. H., Olofsen, E., Aarts, L., Smith, T., Velzen, M. van, et al.: ‘Opioid utility function: methods and implications’; Annals of Palliative Medicine, 9, 2 (2020), 528–536. https://doi.org/10.21037/apm.2019.10.09

[van der Walt et al., 2011] van der Walt, S., Colbert, S. C., Varoquaux, G.: ‘The NumPy Array: A Structure for Efficient Numerical Computation’; Computing in Science & Engineering, 13, 2 (2011), 22–30. https://doi.org/10.1109/MCSE.2011.37

[Wilensky, 1999] Wilensky, U.: ‘NetLogo’; Evanston: Center for Connected Learning and Computer-Based Modeling, Northwestern University (1999). Retrieved from https://ccl.northwestern.edu/netlogo/

[Wooldridge and Jennings, 1995] Wooldridge, M., Jennings, N. R.: ‘Intelligent agents: theory and practice’; The Knowledge Engineering Review, 10, 02 (1995), 115–152. https://doi.org/10.1017/S0269888900008122

[Zhang and Zheng, 2019] Zhang, N., Zheng, X.: ‘Agent-based simulation of consumer purchase behaviour based on quality, price and promotion’; Enterprise Information Systems, 13, 10 (2019), 1427–1441. https://doi.org/10.1080/17517575.2019.1654133

[Zhang and Zhang, 2007] Zhang, T., Zhang, D.: ‘Agent-based simulation of consumer purchase decision-making and the decoy effect’; Journal of Business Research, 60, 8 (2007), 912–922. https://doi.org/10.1016/j.jbusres.2007.02.006