
Article:
Yang, H. orcid.org/0000-0002-3372-4801 and Bath, P.A. orcid.org/0000-0002-6310-7396 (2018) Predicting loneliness in older age using two measures of loneliness. International Journal of Computers and Applications, 42 (6). pp. 602-615. ISSN 1206-212X

https://doi.org/10.1080/1206212X.2018.1562408

Predicting Loneliness in Older Age Using Two Measures of Loneliness

Hui Yang and Peter A. Bath

Information School, University of Sheffield, Sheffield, United Kingdom

Corresponding author:

Dr. Hui Yang

Information School

University of Sheffield

Regent Court, 211 Portobello Street

Sheffield, S1 4DP, United Kingdom

E-mail: hyang@sheffield.ac.uk

Notes on contributors:

Hui Yang, PhD, is currently working as a Research Fellow in the Information School, University of Sheffield, UK. Her area of interest includes health informatics, bioinformatics, business intelligence, and social media.

Peter A. Bath, PhD, is the Professor of Health Informatics in the Information School, University of Sheffield, UK. His research interests include analysing health data using statistical methods and data mining techniques, the evaluation of health information systems and the use of digital information in health care of older people.

Predicting Loneliness in Older Age Using Two Measures of Loneliness

Older people are especially vulnerable to loneliness and this has become a major public health concern for people in later life. In this paper, we propose a machine learning based approach to predict loneliness probability using two gradient boosting algorithms, XGBoost and LightGBM. The predictive models are built using data from a large nationally representative sample from, the English Longitudinal Study of Ageing (ELSA) that had 7 successive waves (2002 ~ 2015). Two measures of loneliness were applied to investigate the impact of different measure strategies on the prediction of loneliness. The models achieved good performance with a high Area Under Curve (AUC) and a low Logarithmic Loss (LogLoss) on the test data, i.e., AUC (0.88) and LogLoss (0.24) using the single-item direct measure of loneliness, and AUC (0.84) and LogLoss (0.31) using the multi-item indirect measure of loneliness. A wide range of variables were investigated to identify significant risk factors associated with loneliness. Specific categories associated with important variables were also recognized by the models. Such information will further enhance our understanding and knowledge of the causes of loneliness in elderly people.

Keywords: loneliness; measure of loneliness; older age; ELSA data; predictive model; gradient tree boosting

Subject classification codes: Applied computing; Health care information systems

1. Introduction

Loneliness can be understood as subjective measure of unwelcome feelings or perceptions on the part of the respondent, associated with lack an affection, closeness, and social connection with others [27]. Older people are especially vulnerable to loneliness: according to recent figures from Age UK, over 1 million people aged over 60 in England say they are always or often feel lonely.

Loneliness is a major public health concern for people in later life. Research [3] has shown that loneliness could have a profound impact on the deterioration of physical

and mental health and can reduce quality of life. When loneliness becomes persistent, it may even lead to an early death [19].

Several studies have shown that loneliness is not a permanent condition but rather can be a transient, recurrent, or persistent (chronic) state [38, 39]. Loneliness might be affected by changes in environmental or living conditions, for example, major life changes (e.g., retirement, bereavement, etc.), which can occur in older age, and these can trigger the feelings of loneliness either temporarily or over a longer term.

Many existing studies have found a wide range of risk factors to be correlated with older people indicating they feel lonely. Some research [11, 12, 17, 37] has shown a strong association between loneliness and socio-demographic factors such as gender, age, ethnicity, education, marital status, and living arrangements. There is evidence that social isolation (e.g., the lack of contact with family or friends, community involvement, or access to services) plays a key role in the reporting of feelings of loneliness [10, 18, 39, 23, 24, 28]. Moreover, poor physical health, long-standing illness, reduced mobility, mental health like cognitive impairment, and sensory impairment increase older people's chances of being lonely [6, 7, 25, 30, 32, 35, 36], as they can reduce people's capacity to get out and about and develop and maintain friendships. Some research has highlighted that there is a direct correlation between economic status and loneliness among old people [15], i.e., poorer individuals are more likely to experience loneliness than those who are financially well off.

Loneliness is a complex and multidimensional construct and measuring loneliness is a subjective experience and can be complicated. Direct questions, such as '*do you feel lonely*?' can lead to an under-representation in the dataset, as some respondents may not wish to admit to experiencing loneliness. Negatively worded statements such as '*I often feel isolated from others*' may also cause some respondents to misrepresent their loneliness and result in underestimates in the occurrence of

loneliness. Different measures might be more or less appropriate in different settings [34].

Current studies on loneliness have four main limitations: first, while most previous studies have focused on the research about association of risk factors with loneliness in older adults within a single measurement wave, there is limited robust research on predicting the likelihood of loneliness from a longitudinal perspective. Second, loneliness is a complicated construct, which is associated with a wide range of factors, and sensitive to changes in environmental conditions. However, previous research has mostly investigated the impact of risk factors on loneliness from one aspect of loneliness, and little work has considered a wide variety of variables reflecting complete and comprehensive aspects of loneliness; nor has it taken into account the impact of variables from previous waves into account when building predictive models. Third, most existing research has used descriptive statistics and logistic regression models for studying loneliness from a social science perspective. Although this work has been valuable in developing better insights into loneliness, there has been limited research utilising advanced data mining methods: here we seek to fill this gap in research by using advanced machine learning approaches. Fourth, although various measures of loneliness have been developed, there has been little research that has compared the correlates of different measures of loneliness. It would be advantageous for studies to examine the relative strengths of different measures associated with the these measures, or feelings, of loneliness.

In this paper, two measures of loneliness were used for the study, one is a single-item direct measure, and the other is a multi-item indirect measure. Predictive models with respect to these different measures were developed separately to determine whether a person is at an increased risk of experiencing lonely. The models were built on a nationally representative sample from the English Longitudinal Study of Ageing

(ELSA)[1], which contained seven successive waves of data collected during the period from 2002 to 2015. A wide variety of variables were extracted from both previous and current waves. Two gradient boosting based algorithms, XGBoost [5] and LightGBM [22], were used to implement separate predictive models for comparing the model performance. Two evaluation metrics, AUC (Area Under the roc Curve) [33] and LogLoss (Logarithmic Loss) [29], were used to estimate the effectiveness of the predictive models. A list of important risk factors that had significant impacts on loneliness were identified using these models. The model performance and the identified risk factors for the two loneliness measures were compared and explored to uncover possible potential correlations between a range of factors and these two measures.

The overall contribution of this study is that it addresses the issue of loneliness in older adults from a new perspective, i.e., using data mining methods for predicting loneliness in older people. Specifically, we first explore the influence of different measures on the prediction of loneliness using data from a large nationally-representative study of older people, and then develop predictive models by combining two gradient boosting based algorithms with a wide variety of variables to improve the performance. Finally, we identify separate sets of risk factors for the two loneliness measures and compare those identified in the two feature lists. Identifying such information is potentially useful in the design of interventions to prevent or alleviate loneliness in older people.

[1] https://www.elsa-project.ac.uk

2. Methods

2.1. Gradient Boosting Machine (GBM)

Gradient tree boosting [14] is also known as gradient boosting machine (GBM) or gradient boosted regression tree (GBRT). It is extremely powerful, and has been shown to outperform other well-established machine learning methods, such as Hidden Markov Model [4], Neural Networks [20], and Support Vector Machine (SVM) [8], for example, by winning data science competitions such as Kaggle[2]. For a given data set with *n* examples and *m* features, a tree ensemble model uses *K* additive functions to predict the output:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}, \qquad (1)$$

where $\mathcal{F} = \{f(x) = w_{q(x)}\}(q: \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ is the space of regression trees. Here $q(x)$ is the function that maps an example x to the corresponding leaf index in the structure of the tree. Each $f(x)$ corresponds to an independent tree structure q and a leaf weight w. T is the number of path (leaves) in the tree. A path is ended with a leaf that contains weight w. Given an example, we use decision rules in the trees (given by q) to classify it into the leaves and calculate the final prediction by summing up the score in the corresponding leaves (given by w) [14].

To learn the set of functions used in the model, we minimize the regularized objective below:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \qquad (2)$$

---

[2] https://www.kaggle.com

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||w||^2 \qquad (3)$$

where $\mathcal{L}(\phi)$ is a loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i. $\Omega(f_k)$ is a regularization term that helps smooth the final learnt weights to avoid over-fitting. For further details about gradient tree boosting, please refer to the paper by Friedman [14]

2.2. XGBoost vs. LightGBM

Both XGBoost (eXtreme Gradient Boosting) [5] and LightGBM (Light Gradient Boosting Machine) [22] are distributed, high-performance gradient boosting frameworks based on decision tree algorithms. They are well recognized for ranking, classification, and many other machine learning tasks. Although both algorithms are decision tree based, they still vary in a few specific ways, especially in how they create decision trees. LightGBM uses leaf-wise splitting rather than depth-wise splitting, which enables it to converge much faster but also leads to overfitting. XGBoost uses a more regularized model formalization to control overfitting, which gives it better performance. Unlike XGBoost, which is a relatively mature toolkit and is widely utilized by many real-world applications, one limitation for LightGBM in its usage is because it has been developed only recently, and there is less documentation available.

2.3. Parameter tuning

Unlike other machine learning algorithms that have no or a few parameters to be tuned, Gradient Boosting based algorithms have one disadvantage in that a large number of parameters are needed to adjust in order to obtain the optimal performance. In an ideal world, with infinite resources and where time is not an issue, it is possible to run a giant search, with all the parameters together, and find the optimal solution. However, in the real world, the parameter turning process becomes computationally more expensive as

the dataset grows bigger, and the training time grows as well. For this reason, it is important to understand the role of parameters and focus on the steps that we expect to impact our results the most. Table 1 lists a number of XGBoost or LightGBM hyperparameters that usually have an important impact on performance in terms of faster speed, better accuracy, or overfitting prevention.

[Table 1 is here]

In our predictive models, both algorithms adopt a hyperparameter optimization method, called the grid search technique, to determine the optimal parameters. Grid search [2] is a technique that allows working through multiple combinations of parameter tuning, cross-validating each and determining which one gives the best performance. In addition, both algorithms provide an inherent early-stopping function that is a regularization approach to training complex machine learning models to avoid overfitting.

3. Materials

3.1. ELSA data

The English Longitudinal Study of Ageing (ELSA) is a longitudinal study that collects multidisciplinary data from a representative sample of the population of England aged 50 and older. Fieldwork for the first wave of ELSA was carried out in 2002/2003 with follow-up waves every two years. The present analysis included seven completed ELSA waves (2002 - 2015) and in total this included 17,861 respondents. However, not all the respondents participated all the seven wave surveys, and only 4,088 (22.8%) respondents had provided data at every wave. The distribution of respondents in terms of wave number that they ever participated in is shown in Figure 1.

[Figure 1 is here]

3.2. Dependent variable - loneliness

The ELSA surveys utilized two measures of loneliness as described below:

(1) Single-item measure of loneliness (direct)

The item on loneliness was the question "*How much of the time during the last week, have you felt lonely?*". A single-item measure was produced with the provided dichotomous answer (*no*=0; *yes*=1). The information relevant to the single-item measure of loneliness was available across all the ELSA waves from Wave-1 (2002-03) to Wave-7 (2014-15).

(2) Multi-item measure of loneliness (indirect)

Loneliness was measured by the 3-item short form of the revised UCLA Loneliness Scale [21]. The scale includes indirect questions about *feeling lack of companionship*, *feeling left out* and *feeling isolated from others*. The three-point response scale ranged from 1 (*hardly ever/never*) to 3 (*often*). Ratings were summed to produce a loneliness score ranging from 3 to 9, with a higher score indicating greater loneliness. For the purpose of generating a dichotomous variable, we defined a threshold score ($\geqslant$6) as being lonely in accordance with the work of [35]. The UCLA-based loneliness measure was utilised only from Wave-2 (2004-05) to Wave-7 (2014-15) since the information on the 3-item UCLA loneliness measure was not collected in the wave-1 survey.

3.3. Independent variables

A wide range of variables were generated based on the ELSA data for our loneliness analysis, and these could be categorized into three main groups:

- Baseline variables (28)

The group of variables in Table 2 were treated as baseline, which covers different aspects of information, e.g., socio-demographic status, financial situation, general health condition, and personal behaviour and habit. Baseline variables known to be associated with demographic and economic status are `age`, `gender`, `ethnicity` (white or not), `religion` (church member or not), `marital status` (single, married, separated/divorced, and widowed), `education` (university degree or equivalent, less than university degree, or no qualification), `people number in household`, `retirement on pension` (full/half/no pension), and `money shortage` (often/sometime/not often/never). Social engagement and contact variables include `closeness to spouse` (very close/close/not close), `contact with children and friends` (weekly/monthly/yearly or rare contact). Health-related variables contain some self-reported variables like `general health`, `long-standing illness`, `disability that limits work`, `eyesight`, and `hearing`. Psychological (e.g., `depression`) and health behaviour variables (e.g., `alcohol`, `smoking`, and `sport activity`) were also considered in the study.

[Table 2 is here]

- Disease-related variables (19)

The variable set in Table 3 mainly contains various diseases closely associated with elderly people, such as various circulatory diseases (e.g., `high blood pressure`, `angina`, `heart attack`, `heart failure`, and `heart murmur`), long-standing illnesses (e.g., `diabetes`, `cancer`, `stroke`, `arthritis`, and `lung disease`), and diseases relevant to cognitive impairment and psychiatric problems (e.g., `Alzheimer's disease`, `dementia`, and `psychiatric condition`). It should be noted that most of the (binary) disease variables had a

high level of missing data (i.e., over 90%). A possible explanation for this is that the participants might have had no such disease or were not willing to disclose details about their health conditions.

[Table 3 is here]

- Disability-related variables (28)

Table 4 demonstrates that disabilities were assessed based on the participants responses to interviewers' questions on perceived difficulties in six basic activities of daily living (ADLs), such as difficulty in `dressing`, and seven instrumental ADLs (IADLs), such as `preparing a hot meal`. Participants with difficulties in one or more of these activities were considered to have some degree of disability. Mobility impairment was defined by asking respondents whether they had difficulties with one or more of 10 common leg and arm functions (e.g., `walking 100 yards`).

An older person with more difficulties in ADLs, IADLs, or mobility impairment suggests that they have a higher degree of physical disability. To better reflect the level of disability, summative variables (e.g., `mobility_num`, `IADL_num`, and `ADL_num`) were derived from the associated item variables in individual disability groups by simply summing up all the associated binary items. Higher score indicates greater degree of disability.

[Table 4 is here]

In the selected variables, the majority of the variables were binary or categorical variables, each of which had been encoded with numerical values in the original ELSA data source, e.g., `gender` (1:Famale; 0:Male) and `closeness to spouse` (1:very close;

2: quite close; 3: not close).

It can also be seen in Tables 2-4 that quite a number of variables had missing values with different rate of missing data. For example, some disease-related variables (e.g., `heart attack`, `stroke`) had a high missing rate of above 90%, whereas the disability-related variables had just a few instances of missing values. In this study, we adopted a simple imputation strategy, i.e., to set the missing data with the default value of -1. The reason for us using this strategy was because, according to our experimental results, it outperformed other more complicated strategies such as logistic regression or the most frequent category for binary and categorical variables, and linear regression or mean/median/mode imputation for continuous variables.

3.4. Training, validation, and test data

The dataset studied here contained a total of 69,478 instances, in which each instance corresponds to the value of a variable for a participant in one specific wave. The distribution of participants in different waves is illustrated in Figure 2, and the loneliness prevalence rates, as measured separately by the two loneliness scales, are also plotted in Figure 3. In each wave, at least 9,000 respondents took part in the ELSA survey. For the single-item measure of loneliness, the loneliness prevalence rate of aged adults was relatively stable throughout seven successive ELSA waves, fluctuating within a range of 10% to 14%. The multi-item measure of loneliness led to higher levels of reported loneliness in each wave, and the prevalence rate ranged between 17% - 21% across six consecutive waves (Note that the information about the multi-item measure was missing in wave-1).

[Figure 2 and 3 are here]

As shown in Figure 3, the instances of loneliness are much lower than that for non-loneliness. To deal with the imbalance data issue in the predictive models, both

XGBoost and LightGBM algorithm provide some related parameters, e.g., the parameters `min_child_weight` and `scale_pos_weight` for the XGBoost, and the parameters `scale_pos_weight` and `is_unbalance` for the LightGBM. These parameters were set and tuned optimally to achieve the best performance of the model.

Here we split the whole data into three subsets, training data, validation data and test data sets for model training and prediction. The test data set was generated from the most recent wave, wave-7 (2014/15), the validation data came from wave-6 (2012/13), and the remaining waves were used as the training data. To test the validity of our selected validation data, we examined whether the model performance in the test data was consistent with that of the validation data.

4. Experiments

4.1. Evaluation metrics

The prediction of loneliness can be treated as a binary classification task in which the output is labelled as 1 (loneliness), or 0 (non-loneliness). Given a new instance, the classifier will assign a loneliness probability to the instance rather than simply yielding the most likely class label. The predicted probability will fall in the range of [0, 1]. A higher predicted probability means that a participant is more likely to be lonely.

Here two commonly-used evaluation metrics, LogLoss [29] and AUC [33] were applied to estimate model performance. LogLoss (Logarithmic loss) is a classification loss function, which quantifies the accuracy of a classifier by penalizing false classification.

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i] \quad (4)$$

where n is the number of instances. $\hat{y}_i$ is the predicted probability, and y_i is the target class label. A *perfect* classifier should have a LogLoss with the value of *zero*. Less ideal

classifiers have progressively larger values of LogLoss.

AUC (Area Under the ROC Curve) [13] is calculated based on a ROC (Receiver Operating Characteristics) curve that is a plot of the true positive rate against the false positive rate for different possible cut points of predictions with respect to a set of instances. The AUC value lies between 0.5 to 1 where 0.5 denotes a *poor* classifier, i.e., no better than random, and 1 denotes a *perfect* classifier.

The two evaluation metrics used differ in several ways when used for binary classification problems: (1) What LogLoss likes is that predicted probability belonging to one class is as close as possible to its true probability at individual instance level. On the contrary, for AUC, the predicted probability for one instance does not matter. AUC improves when the order of the predictions corresponding to a set of instances becomes more correct. (2) For LogLoss, the lower the score is, the better the model performs. For AUC score, the higher the score the better the model performs.

4.2. Experiment setup

To compare the impact of different sets of variables on loneliness, here we developed several predictive models:

- **Model I** (Baseline variables in the current wave)
- **Model II** (Disease-related variables in the current wave)
- **Model III** (Disability-related variables in the current wave)
- **Model IV** (Full set of variables in the current wave: Baseline + Disease + Disability)
- **Model V** (Full set of variables in both the previous and the current wave plus the loneliness value in the previous wave)

In Model V, additional information from the previous wave was considered and more variables from the previous wave were extracted and then added into the models for the analysis. For example, in the test data used for Model V, there were two different variables related to the information of the disease `Diabetes`. One was $\texttt{Diabetes}_{\texttt{curr}}$ from the current wave (wave-7), and the other was $\texttt{Diabetes}_{\texttt{prev}}$ from the previous wave (wave-6). These two variables were probably different as an older adult who had no diabetes in the wave-6 (2012/13) might develop this disease in the wave-7 (2014/15). It should be noted that a few special variables (e.g., `gender`, `education`) in the previous wave were excluded from Model V because there was no change between the two waves with respect to these variables. For example, only $\texttt{gender}_{\texttt{curr}}$ rather than $\texttt{gender}_{\texttt{prev}}$ was used in Model V. In addition, the dependent variable in the previous wave, $\texttt{loneliness}_{\texttt{prev}}$, was also added as an important feature into Model V. Hence, the feature set used in Model V in practice consists of three parts, the independent variables in the current wave, independent variables in the previous wave, and the variable $\texttt{loneliness}_{\texttt{prev}}$. The outcome variable of Model V was $\texttt{loneliness}_{\texttt{curr}}$.

To examine possible differences in the model performance and risk factors of importance identified by different measures of loneliness, two sets of experiments were conducted, each of which was targeted for one particular type of measure.

5. Results

5.1. Model performance comparison at the variable level

5.1.1. Performance comparison in different machine learning algorithms

To evaluate the performance of gradient boosting algorithms, we used the commonly-used algorithm, Logistic Regression, as a baseline. Three selected ML algorithms performed consistently for both measures of loneliness. As shown in Table 5, both gradient boosting algorithms, XGBoost and LightGBM, generally performed better than

Logistic Regression with an improvement between 1-5 percent in both AUC and LogLoss. Furthermore, both XGBoost and LightGBM achieved competitive performance on both the validation and test data in terms of AUC and LogLoss.

[Table 5 is here]

Table 6 provides the total running time (second) spent at the training, validation, and test stage on the same computer (iMac Desktop, 2.7GHz Intel Core i5 CPU, and 32G RAM). It is obvious that Logistic Regression ran the fastest due to its relatively simple learning process. As for work efficiency, LightGBM took much less time than XGBoost but it achieved the similar performance to XGBoost.

[Table 6 is here]

5.1.2. Performance comparison between different groups of variables

A series of experiments (see Table 5) was conducted to compare model performance using different sets of variables (see subsection 4.2 – Experiment setup). As described above, we grouped the variables into three functional subsets: baseline variables (Model I) performed the best, followed by disability-related variables (Model II), and disease-related variables (Model III) the worst. It is noted that when all the variables in the current wave were combined, the performance of Model IV was slightly improved, compared with the baseline model (Model I). We the evaluated these results using the Wilcoxon signed-rank test. The p-value from the Wilcoxon signed-rank test shown in Table 7 indicates that the difference between these two models in terms of AUC and Logloss was significant. It implies that the addition of the disease-related and disability-related variables significantly improved the prediction of loneliness. It is also interesting to note that when the information in the previous wave was added into Model V, model accuracy was significantly increased compared with Model IV (see Table 7). This suggests that some variables might have a persistent influence on loneliness, and the

context is needed to track back to a relatively longer period.

[Table 7 is here]

5.1.3. Performance comparison between validation data and test data

It is noted that, in Table 2, the performance of the predictive models using the three ML algorithms was quite stable and there was no dramatic change in model performance between the validation data set and the test data set. This implies that the test data display similar characteristics to the validation data, and the data-splitting strategy was appropriate.

In the single-item measure of loneliness, the predictive models generally performed better on the validation data than on the test data for the AUC but not for the LogLoss. In the multi-item measure of loneliness, the performance on the test data was slightly better than that of the validation data in terms of both AUC and Logloss.

5.1.4. Risk factor comparison between different GBM algorithms (Model IV)

Figure 4 provides the relationship between the Logloss error and the number of the top-ranking features in terms of different ML algorithms and loneliness measures. The Logloss was drastically reduced until the number of the top-ranking features approached 30, then it tended to be stable after this threshold point. This implies that the top-30 ranking features have an important impact on the prediction of loneliness. Due to the limit of the paper space, we here select the top-20 ranking features for feature comparison in terms of different ML algorithms and loneliness measures.

[Figure 4 is here]

Table 8 provides the top-ranking predictive variables with high importance scores in Model IV (Full data in current wave) regarding the two measures of loneliness. Due to the difference in implementation by XGBoost and LightGBM, the

top-ranking lists generated by both algorithms were different to some extent, even within the same measure of loneliness.

[Table 8 is here]

In the single-item measure, 12 out of the top 20 risk factors were identified by both XGBoost and LightGBM. XGBoost identified more disease-related and disability-related variables as being important, e.g., disease-related variables (`Psychiatric condition`, `Arthritis`, `Asthma`, and `High blood pressure`) and disability-related variables (`Lifting weights over 10 pounds` and `Sitting for 2 hours`), whereas LightGBM identified more variables related to social factors (e.g., `Education`, `People number in household`), Health Behaviour (e.g., `Alcohol: how many days per week`), and Sensory Impairment (e.g., `Self-reported hearing` and `Self-reported eyesight`).

In the multi-item measure, as many as 17 variables in the top-20 ranking list were recognised by both GBM algorithms. The high overlap rate between these two important risk factor lists suggests that the identification of loneliness-related risk factors has a more stable and consistent performance using the multi-item measure method.

5.1.5. Risk factor comparison between different measures of loneliness (Model IV)

It is interesting to examine whether the predictive models built by the same ML algorithm could identify different risk factors of importance when employing different measures of loneliness.

It can be seen from Table 3 that, for the XGBoost algorithm, the top-ranking lists in both measures were quite different. Nearly a half of the risk factors in the single-item measure did not appear in the list of the multi-item measure. For the LightGBM, the risk factors identified by both measures were almost identical, although the

importance ranking order was changed. This suggests that the XGBoost is more sensitive to the selection of the loneliness measure than the LightGBM.

5.1.6. Risk factor comparison in Model V

To find out which variables in the previous wave are helpful in the prediction of loneliness, the top-ranking risk factors generated by Model V (Full data in both the previous and the current wave) are listed in Table 9. It can be seen that, when the information in the previous wave was added into the predictive model, several variables in the previous wave showed a statistically significant association with loneliness, which included: `Loneliness`, `Marital status`, `Age`, `Closeness to spouse`, `Contact with friends`, `Depression`, `Self-reported health`, and disability-related measures (`Mobility_num` and `IADL_num`). At the same time, such variables are also recognised as important in the current wave. It implies that these variables might have more long-term association with feelings of loneliness.

[Table 9 is here]

The identical risk factors identified by both XGBoost and LightGBM were 15 in the single-item measure and 14 in the multi-item measure individually. The high overlap of risk factors between the two ML algorithm means that most of the risk factors are important, and thus are less sensitive to the selection of the ML algorithms.

When different measures of loneliness were applied, only a few risk factors (4 - 6) were different in the top-20 ranking lists created by the same ML algorithm. This suggests, to some extent at least, that there exists potential correlation between the risk factors for these two measures of loneliness.

5.1.7. Risk factor comparison between different models

Table 10 presents the top-10 ranking features in different XGB-based models in terms of different loneliness measures. For both loneliness measures, there is a high level of

overlap between the top-ranking features in both Model I and II. For Model III, three variables obviously outperform other individual item variables, and are located in the top-3 list. It should be noted that there is just slight change between the top-ranking features in Model I (baseline variables) and Model IV (full variables) with 1-2 new variables appearing in Model IV.

[Table 10 is here]

5.2. Model performance comparison at the variable category level

As discussed earlier, a subset of variables were recognized as key risk factors for loneliness. However, as we know, most of the variables are categorical variables, each of which has several categories. For example, `Marital Status` is divided into 6 different categories, i.e. 'single' (never married), 'married' (firstly and only married), 'remarried' (second or later marriage), 'legally separated', 'divorced', and 'widowed'. `Closeness to spouse` is coded as 4 levels, i.e. 'very close', 'quite close', 'not very close' and 'not at all close'. It is assumed that not all the categories within one variable are of equal importance for the prediction of loneliness. It would be interesting to find which categories are more important than others.

To conduct a more fine-grained analysis at the variable category level, each categorical variable was first converted into dummy variables using one-hot encoding. A sparse matrix with numerous dummy variables was created to build a new set of predictive models.

[Table 11 is here]

Table 11 presents the performance of the models with different variable sets at the variable category level. Compared with the performance at the variable level (recall Table 2), it appears that, following the dummification of the categorical variables, the model performed similarly to the one at the variable level where categorical variables

were only encoded as numeric values. This implies that the additional dummy variable information did not help improve the model performance. Nevertheless, the best performance was still achieved when the information from both previous and current waves (Model V) were combined together.

[Table 12 and 13 are here]

Tables 12-13 list the top-20 variable categories in Model IV (Full data in current wave) and Model V (Full data in both previous and current wave). Some specific categories associated with important variables in both the previous and the current wave were identified by the models. For instance, 'married' and 'widowed' from `Marital Status`, 'very close' and 'quite close' from `Closeness to spouse`, '1-2 time a week' and '>=3 times a week' from `Contact with friend` and `Contact with children`, 'good' from `Self-reported eyesight`, 'often' from `Health problem`, and 'often' from `Money shortage`, were more likely to be associated with loneliness.

It is observed that missing data (N/A) from some variables were also recognized as key predictive factors by the models. As we know, part of the missing data may be due to the respondents' reluctance to disclose their true feelings to some sensitive questions such as the relationship with their spouse, connections with children and friends. In our studies, the missing data in the categorical variables were replaced with -1, and thus were treated as one special category.

6. Discussion

Loneliness can have profound impact on health and well-being, and thus it has become one of the leading concerns for well-being among older people. In this paper, we investigated the association of loneliness with a wide variety of factors extracted from seven ELSA waves conducted from 2002 to 2015. The results of our study support the findings of several studies [1, 31, 35] that have suggested strong connections between

loneliness and age, gender, and living arrangements. There is clear evidence that the likelihood of expressing feelings of loneliness increases with age and that men and women are affected differently by loneliness: older women are more likely to report feeling lonely than older men. Moreover, people who live on their own are more prone to experiencing loneliness than those living with others.

Our research has shown that the extent of the emotional attachments between respondents and their family member and friends play an important role in the prevalence of loneliness. Married people who live in an emotionally close relationship with their spouse report lower levels of loneliness. As widowhood leads to the end of a positive attachment, it is not surprising that higher levels of emotional loneliness are experienced following the death of a spouse. Furthermore, evidence shows that frequent social contact with children and friends can lower the level of social isolation and loneliness. These findings are consistent with the multi-variate analyses by Dahlberg et al. [11] and the analyses by Heikkinen and Kauppinen [16].

In the categorical variable `Money shortage`, three out of four categories, `often, sometimes,` and `never`, co-occur in the top-ranking variable list (see Table 11). It suggests that `Money shortage` is an important risk factor for loneliness. In the variable `Education`, `no qualification` was recognised as the top-ranking category compared with other levels of education. Some categorical variables related to lifestyle and behaviour such as `Alcohol[No]`, and participating in less sport or exercise, `Mild sport/activity[>1 times a week]` and `Vigorous sport/activity[Hardly /never]`, were seen as being influential on loneliness. This accords with the findings by Luo et al. [25].

In line with previous research [1, 26], the most important difference between Model I and Model IV was reflected in the Wilcoxon signed-rank test (Table 7), which confirmed the strong association between various health problems and loneliness. Some

physical illnesses (e.g., `Lung disease`, `Arthritis`, `Asthma`), heart-related diseases (e.g., `High blood pressure`), and mental health problems (e.g., `Depression` and `Psychiatric condition`) were identified as important risk factors in the top-ranking feature list (see Table 8). Some difficulties with mobility (e.g., `sitting for 2 hours`) and maintaining daily activities (e.g., `Lifting weights over 10 pounds`) might lead to neglect and social isolation associated with loneliness.

Further evidence from our study (Table 9) suggests that some variables (e.g., `Age`, `Depression`, `Marital status`, and `Closeness to spouse`) have a relatively long-term influence on the feelings of loneliness. Hence, it would be more meaningful to consider some historical data (e.g., the information from the previous wave) in predicting loneliness from a longitudinal viewpoint.

The analyses on the prediction results using Model V shown in Table 14 indicate that the XGBoost and LightGBM models are highly correlated as the correlation score using the Pearson method is over 0.95 on the single-item measure and above 0.88 on the multi-item measure in the prediction of the test data (wave-7). A possible explanation for this is that both algorithms had similar lists of important risk factors, although the ordering of the risk factors was slightly different.

[Table 14 is here]

The evidence from the results also indicated that the performance of the predictive models might be affected by the selection of loneliness measures. The results of the Chi-squared tests in Table 15 show that the single-item direct measure was significantly associated with the multi-item indirect measure across different waves. The strong association between these two measures might, at least in part, explain the high overlap between the top-ranking risk factors of the two measures (see Tables 8-9 and 12-13).

[Table 15 is here]

7. Conclusions

Unlike most existing studies that work on the loneliness problem in older people using descriptive statistics and logistic regression in social science, this paper has proposed an approach to address from a data science perspective, i.e., using data mining methods. This is a novel approach to predicting loneliness in older age and to identifying risk factors for loneliness in later life: it was undertaken using data from a large nationally-representative sample of older people (ELSA). We developed machine learning based models to predict the likelihood of an older adult being lonely using two gradient boosting algorithms, XGBoost and LightGBM. A large nationally representative dataset from the English Longitudinal Study of Ageing (ELSA) with seven successive waves, was used for the building of loneliness prediction models.

The models achieved good performance with a high AUC and a low LogLoss on the test data, that is, AUC (0.88) and LogLoss (0.24) for the single-item measure of loneliness, and AUC (0.86) and LogLoss (0.31) for the multi-item measure of loneliness. Risk factors of significant importance were identified from a wide variety of variables by the predictive models. Specific categories associated with important variable were also recognized by the models, which would help deepen our understanding and knowledge of loneliness causes.

Moreover, two measures of loneliness were applied in this study. The impact of different measures on model performance of predictive models and the identification of important risk factors was also investigated. The results from the study show that the selection of an appropriate measure of loneliness plays an important role for the study of loneliness in older age.

To the best of our knowledge this is the first attempt to develop these types of models for predicting loneliness and risk factors: however, there is still room for improving the accuracy of prediction by exploring more potential risk factors that may

have an influence on the levels of loneliness, e.g., more variables from different domains such as socio-demographics and economy, social engagement and social networks, physical and mental health, and clinical data. More research is also needed for the handling of imbalanced data. Some advanced over-sampling or under-sampling methods like SMOTE (Synthetic minority oversampling technique) will be investigated in future work. Finally, more advanced machine learning algorithms, such as deep learning methods, could be applied in the predictive models, and this might help provide more important insights into the factors affecting loneliness in older age.

Disclosure statement

No potential conflict of interest was reported by the authors.

Acknowledgement

We would firstly like to thank the team of researchers based at University College London, the Institute of Fiscal Studies and the National Centre for Social Research for providing the English Longitudinal Study of Ageing (ELSA) data for free download. Moreover, we also wish to acknowledge the insightful comments and suggestions received from the anonymous reviewers which helped to improve this paper.

References

[1] Aartsen M, Jylhä M. Onset of loneliness in older adults: Results of a 28 year prospective study. Euro J Ageing. 2011;8(1):31–38.

[2] Bergstra JS, Bardenet R, Bengio Y, et al. & Kégl, Algorithms for hyper-parameter optimization. In Advances in neural information processing systems; 2011; p.2546-2554.

[3] Bennett KM. Low level social engagement as a precursor of mortality among people in later life. Age and Ageing. 2002;31(3):165–168.

[4] Baum, LE, Petrie T. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics. 1966;37(6):1554–1563.

[5] Chen T, He T. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on knowledge discovery and data mining; 2016; p.785–794.

[6] Christiansen J, Larsen, FB, Lasgaard M. Do stress, health behaviour, and sleep mediate the association between loneliness and adverse health conditions among older people? Social Science & Medicine. 2016;152:80-86.

[7] Cole A. Mental health and loneliness in old age. Nursing standard. 2016;30(32):20-21.

[8] Cortes C, Vapnik VN. Support-vector networks. Machine Learning. 1995;20(3):273–297.

[9] Dahlberg L, Andersson L, McKee, KJ, et al. Predictors of loneliness among older women and men in Sweden: A national longitudinal study. Aging Ment Health. 2015;19(5):409–417.

[10] Dahlberg L, Andersson L, Lennartsson C. Long-term predictors of loneliness in old age: results of a 20-year national study. Aging & Mental Health. 2018;22(2):190-196.

[11] de Jong Gierveld J, Broese van Groenou M, Hoogendoorn AW, et al. Quality of Marriages in Later Life and Emotional and Social Loneliness. J Gerontol Ser B, Psychol Sci and Soc Sci 64B. 2009;4:497–506.

[12] Dykstra PA. Older adult loneliness: Myths and realities. Euro J Ageing. 2009;6(2):91–100.

[13] Freeman EA, Moisen GG. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecological Modelling. 2008; 217(1):48-58.

[14] Friedman J. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189-1232.

[15] Fokkema T, de Jong Gierveld J, Dykstra, PA. Cross-National Differences in Older Adult Loneliness. J Psychol. 2012:146(1-2):201–228.

[16] Heikkinen RL, Kauppinen M. Mental well-being: a 16-year follow-up among older residents in Jyväskylä. Archiv Gerontol and Geriatr. 2011;52(1):33–39.

[17] Heylen L. The older, the lonelier? Risk factors for social loneliness in old age. Ageing Soc. 2010;30:1177–1196.

[18] Holmen K, Furukawa H. Loneliness, health, and social network among elderly people – a follow-up study. Archiv Gerontol and Geriatr. 2002:35(3):261-274.

[19] Holt-Lunstad J, Smith TB, Baker M, et al. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. Perspect Psychol Sci. 2015;10(2):227-237.

[20] Huang W, Huang R. Network Net and Conventional Classifier. Proceedings of IEEE Conference on Neural Information Processing System-Natural and Synthetic; 1987; p.387-396.

[21] Hughes ME, Waite LJ, Hawkley LC, et al. A short scale for measuring loneliness in large surveys: Results from two population-based studies. Res Aging. 2004;26(6):655–672.

[22] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems; 2017; p.3149–3157.

[23] Landeiro F, Barrows P, Nuttall Musson E, Gray AM, Leal J. Reducing social isolation and loneliness in older people: a systematic review protocol. BMJ Open. 2017;7(5): e013778.

[24] Liu X, Evandrou M, Falkingham J, Feng Z, Vlachantoni A. The Dynamics of Social Support and the Impact on Loneliness among Older People in China. The Gerontologist. 2016;56(Suppl3):436-437.

[25] Luo Y, Hawkley LC, Waite LJ, et al. Loneliness, health, and mortality in old age: A national longitudinal study. Soc Sci Med. 2012;74(6):907-914.

[26] O'Luanaigh C, Lawlor BA. Loneliness and the health of older people. Int J Geriatr Psychiatry. 2008;23:1213-1221.

[27] Peplau, LA, Perlman, D. Perspectives on loneliness. Loneliness: A Sourcebook of Current Theory, Research and Therapy. New York Willey. 1982; p.571-581.

[28] Poscia A, Stojanovic J, La Milia DI. Interventions targeting loneliness and social isolation among the older people: An update systematic review. Experimental Gerontology. 2018;102:133-144.

[29] Rosasco L, De Vito ED, Caponnetto A, et al. Are Loss Functions All the Same? Neural Computation. 2004;16(5):1063–1076.

[30] Shankar A, McMunn A, Banks J, et al. Loneliness, social isolation, and behavioural and biological health indicators in older adults. Health Psycho. 2011;30(4):377-385.

[31] Shankar A, Rafnsson SB, Steptoe A. Longitudinal associations between social connections and subjective wellbeing in the ELSA. Psychol Health. 2015;30(6):686-698.

[32] Smith Kj, Victor C. Investigating the relationship between diabetes and loneliness in older people. Diabetic Med. 2016;33:171.

[33] Sokolova M, & Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag, 2009;45(4):427-437.

[34] Solano CH. Two measures of loneliness: a comparison. Psychol Reports. 1980;46:23-28.

[35] Steptoe A, Shankar A, Demakakos P, et al. Social isolation, loneliness, and all-cause mortality in older men and women. Proceedings of the National Academy of Sciences of the United States of America; 2013; p. 5797-5801.

[36] Taube E, Jakobsson U, Midlöv P, Kristensson J. Being in a Bubble: the experience of loneliness among frail older people. 2016;72(3):631-640.

[37] Tiilikainen E, SeppÄNen M. Lost and unfulfilled relationships behind emotional loneliness in old age. Ageing & Society. 2017;37(5):1068-1088.

[38] Victor C. A longitudinal analysis of loneliness among older people in Great Britain. J Psychol. 2012;146(3):313-331.

[39] Wenger GC, Burholt V. Changes in levels of social isolation and loneliness among older people in a rural area: A twenty-year longitudinal study. Can J Aging. 2004;23(2):115–127.

Word count (7393)

Tables

Table 1. Important hyperparameters tuned in the XGBoost and LightGBM based model

XGBoost

max_depth: Maximum depth of a tree
min_child_weight: Minimum sum of instance weight (hessian) needed in a child
eta(learning rate): Step size shrinkage used in update to prevents overfitting
gamma: Minimum loss reduction required to make a further partition on a leaf node of the tree
subsample: Subsample ratio of the training instances
colsample_bytree: Subsample ratio of columns when constructing each tree
scale_pos_weight: weight of labels with positive class

num_boost_round: the number of boosting rounds or trees to build
early_stopping_rounds: stop the training if performance haven't improved for N rounds

LightGBM

num_leaves: max number of leaves in one tree
learing_rate: Step size shrinkage used in update to prevents overfitting
max_depth: Maximum depth of a tree
min_data_in_leaf: minimal number of data in one leaf
max_bin: Number of bucketed bin for feature values
bagging_fraction: the fraction of the data randomly selected without resampling
bagging_freq: frequency for bagging
is_unbalance: if traing data are unbalanced
scale_pos_weight: weight of labels with positive class

num_boost_round: the number of boosting rounds or trees to build
early_stopping_rounds: stop the training if performance haven't improved for N rounds

Table 2. Basic information on the baseline variables

Baseline variables (28)	**Data type**	**Missing data(%)**
Gender	binary	0
Age	continuous	0
Age group	categorical	0
Ethnicity	binary	0.01
Religion	binary	14.9
Education	categorical	0.78
Marital status	categorical	0.01
Closeness to spouse	categorical	35.3
Contact with children	categorical	28.1
Contact with friends	categorical	17.2
People number in household	continuous	0
Retirement on pension	categorical	87.4
Money shortage	categorical	0
Health problem	categorical	11.4
Self-reported general health	categorical	0.03
Self-reported long-standing illness	binary	0.03
Self-reported disability that limits work	binary	0.11
Self-reported eyesight	categorical	0.02
Self-reported hearing	categorical	0.01
Depression	binary	0.02
Smoking	binary	97.8
Stopped smoking	binary	97.9
Alcohol	binary	23.4
Alcohol: how-often	categorical	12.2
Alcohol: how many days per week	continuous	43.0
Vigorous sport/activity	categorical	0.01
Moderate sport/activity	categorical	0.01
Mild sport/activity	categorical	0.01

Table 3. Basic information on the disease-related variables

Disease-related variables (19)	Data type	Missing data(%)
High blood pressure	binary	70.4
Angina diagnosis	binary	94.0
Heart attack diagnosis	binary	95.5
Congestive heart failure	binary	99.6
Heart murmur	binary	97.0
Abnormal heart rhythm	binary	94.6
Diabetes	binary	92.7
Stroke	binary	97.0
High cholesterol	binary	0
Other heart disease	binary	98.3
Lung disease	binary	96.0
Asthma	binary	90.5
Arthritis	binary	69.2
Osteoporosis	binary	94.5
Cancer	binary	96.7
Parkinsons Disease diagnosis	binary	99.5
Psychiatric condition	binary	92.4
Alzheimers disease	binary	99.9
Dementia	binary	99.6

Table 4. Basic information on the disability-related variables

Disability-related variables (28)	**Data type**	**Missing data(%)**
Mobility Impairment (10)		
Walking 100 yards	binary	0.03
Sitting for 2 hours	binary	0.03
Getting up from chair	binary	0.03
Climbing several flights stairs	binary	0.03
Climbing one flight stair	binary	0.03
Stooping, kneeling or crouching	binary	0.03
Extending arm above shoulder	binary	0.03
Pulling/pushing large object	binary	0.03
Lifting weights over 10 pounds	binary	0.03
Picking up 5p coin	binary	0.03
Activities of Daily Life (ADL)(6)		
Dressing	binary	0.02
Walking across a room	binary	0.02
Bathing/showering	binary	0.02
Eating	binary	0.02
Getting in/out of bed	binary	0.02
Using toilet	binary	0.02
Instrumental Activities of Daily Life (IADL)(9)		
Using map in strange place	binary	0.02
Recognising physical danger	binary	0.02
Preparing a hot meal	binary	0.02
Shopping for groceries	binary	0.02
Making phone	binary	0.02
Communication	binary	0.02
Taking medicine	binary	0.02
Doing work around house/garden	binary	0.02
Managing money	binary	0.02
Statistics(3)		
Mobility_num	numeric	0.03
IADL_num	numeric	0.02
ADL_num	numeric	0.02

Table 5. Performance comparison for different predictive models according to number of variables in Models I-V.

AUC (validation/test)						
	Single-item measure of loneliness			**Multi-item measure of loneliness**		
	XGBoost	**LightGBM**	**Logistic Reg.**	**XGBoost**	**LightGBM**	**Logistic Reg.**
Model I	0.858/0.850	0.846/0.835	0.823/0.820	0.808/0.813	0.805/0.809	0.747/0.752
Model II	0.622/0.621	0.612/0.606	0.607/0.606	0.602/0.600	0.600/0.593	0.604/0.602
Model III	0.688/0.667	0.675/0.653	0.675/0.649	0.644/0.639	0.641/0.636	0.641/0.636
Model IV	0.862/0.853	0.853/0.844	0.826/0.824	0.811/0.818	0.808/0.815	0.751/0.757
Model V	0.882/0.887	0.877/0.878	0.857/0.858	0.855/0.875	0.848/0.866	0.815/0.836
Logloss (validation/test)						
	Single-item measure of loneliness			**Multi-item measure of loneliness**		
	XGBoost	**LightGBM**	**Logistic Reg.**	**XGBoost**	**LightGBM**	**Logistic Reg.**
Model I	0.260/0.251	0.265/0.257	0.284/0.269	0.401/0.363	0.404/0.366	0.442/0.399
Model II	0.346/0.332	0.347/0.334	0.359/0.344	0.494/0.455	0.495/0.456	0.495/0.455
Model III	0.334/0.323	0.337/0.326	0.344/0.332	0.485/0.447	0.486/0.448	0.487/0.446
Model IV	0.258/0.249	0.263/0.254	0.283/0.268	0.398/0.360	0.402/0.362	0.439/0.394
Model V	0.243/0.226	0.249/0.232	0.263/0.241	0.356/0.311	0.360/0.314	0.390/0.346

Table 6. The total running time (seconds) for different predictive models according to the number of variables in Models I-V.

AUC (train+validation+test) (second)						
	Single-item measure of loneliness			Multi-item measure of loneliness		
	XGBoost	**LightGBM**	**Logistic Reg.**	**XGBoost**	**LightGBM**	**Logistic Reg.**
Model I	13.118	0.324	0.224	18.430	8.252	0.659
Model II	2.277	0.397	0.222	6.761	0.624	0.199
Model III	3.165	0.769	0.111	3.292	0.790	0.071
Model IV	37.952	2.531	1.318	35.273	12.393	1.349
Model V	47.159	4.037	1.733	35.362	4.978	1.835
Logloss (train+validation+test) (second)						
	Single-item measure of loneliness			Multi-item measure of loneliness		
	XGBoost	**LightGBM**	**Logistic Reg.**	**XGBoost**	**LightGBM**	**Logistic Reg.**
Model I	13.506	5.670	0.223	15.049	8.637	0.666
Model II	10.886	3.753	0.224	7.756	4.494	0.195
Model III	16.896	4.810	0.095	7.206	5.450	0.061
Model IV	34.763	7.914	1.326	38.384	13.400	1.547
Model V	41.245	8.601	0.668	32.363	13.661	1.757

Table 7. Results of the Wilcoxon signed-rank test between different model pairs

	AUC		Logloss	
	Test z-value	P-value	Test z-value	P-value
Model I vs. Model IV	-3.07	0.002	-3.84	0.002
Model IV vs. Model V	-3.065	0.002	-3.065	0.002

Note: The 95% Confidence Interval (CI) level is used.

Table 8. Top-ranking variables in Model IV (full data in current wave) at the variable level

	Single-item measure of loneliness	**Multi-item measure of loneliness**
	XGBoost Model	**XGBoost Model**
1	Depression	Age
2	Closeness to spouse	Contact with friends
3	Marital status	Contact with children
4	Age	Closeness to spouse
5	Mobility_num	Money shortage
6	Money shortage	Education
7	Health problem	Alcohol: how-often
8	Psychiatric condition	Health problem
9	Contact with children	Self-reported hearing
10	Contact with friends	Mobility_num
11	Gender	Marital status
12	IADL_num	Alcohol: how many days per week
13	Lung disease	Self-reported eyesight
14	ADL_num	Self-reported health
15	Vigorous sport/activity	People No. in household
16	Arthritis	Moderate sport/activity
17	Lifting weights over 10 pounds	Gender
18	Asthma	IADL_num
19	Sitting for 2 hours	Mild sport/activity
20	High blood pressure	Religion
	LightGBM Model	**LightGBM Model**
1	Age	Age
2	Closeness to spouse	Contact with friends
3	Mobility_num	Contact with children
4	Marital status	Money shortage
5	Self-reported health	Mobility_num
6	Money shortage	Education
7	Contact with children	Closeness to spouse
8	Psychiatric condition	Health problem
9	Education	Alcohol: how-often
10	Alcohol: how-often	People No. in household
11	People No. in household	Self-reported hearing
12	IADL_num	Marital status
13	Health problem	Alcohol: how many days per week
14	Gender	Self-reported health
15	Self-reported eyesight	Self-reported eyesight
16	Contact with friends	IADL_num
17	ADL_num	Gender
18	Alcohol: how many days per week	ADL_num
19	Moderate sport/activity	Psychiatric condition
20	Self-reported hearing	Moderate sport/activity

Table 9. Top-ranking variables in Model V (full data in both previous and current wave and loneliness in previous wave) at the variable level

	Single-item measure of loneliness	**Multi-item measure of loneliness**
	XGBoost Model	**XGBoost Model**
1	Loneliness (prev.)	Age
2	Depression	Closeness to spouse
3	People No. in household	Contact with friends
4	Closeness to spouse	Money shortage
5	Marital status	Contact with children
6	Closeness to spouse (prev.)	Age (prev.)
7	Marital status (prev.)	Health problem
8	Age (prev.)	Alcohol: how-often
9	Age	Loneliness (prev.)
10	Contact with children	Contact with children (prev.)
11	Self-reported health	Education
12	Depression (prev.)	Mobility_num
13	Alcohol: how-often	Contact with friends (prev.)
14	Money shortage	Self-reported hearing
15	Gender	Marital status
16	Education	Alcohol: how many days per week
17	Mobility_num	People No. in household
18	Contact with friends (prev.)	Self-reported health
19	Self-reported eyesight	Closeness to spouse (prev.)
20	Contact with friends	Self-reported eyesight
	LightGBM Model	**LightGBM Model**
1	Closeness to spouse	Closeness to spouse
2	Age	Age
3	Marital status	Health problem
4	Mobility_num (prev.)	Money shortage
5	People No. in household	Mobility_num
6	Mobility_num	Contact with friends
7	Self-reported health	Mobility_num (prev.)
8	Education	Contact with children
9	Self-reported health (prev.)	Loneliness (prev.)
10	Money shortage	Closeness to spouse (prev.)
11	Marital status (prev.)	IADL_num
12	People number in household (prev.)	Depression
13	Depression	Health problem (prev.)
14	Alcohol: how-often	ADL_num
15	Age (prev.)	Money shortage (prev.)
16	Self-reported eyesight	People No. in household
17	Marital status (prev.)	Alcohol: how many days per week
18	Health problem	Self-reported health (prev.)
19	IADL_num (prev.)	Marital status
20	Contact with children	Self-reported health

Table 10. Top-10 ranked features in different models

	Single-item measure of loneliness	**Multi-item measure of loneliness**
Model I	Age Contact with children Education Contact with friends Self-reported health Self-reported eyesight Money shortage Closeness to spouse Marital status Alcohol: how-often	Age Contact with friends Contact with children Education Alcohol: how-often Money shortage Health problem Closeness to spouse Self-reported health Self-reported hearing
Model II	High blood pressure High cholesterol Angina diagnosis Asthma Osteoporosis Heart attack diagnosis Lung disease Diabetes Arthritis Abnormal heart rhythm	High cholesterol Angina diagnosis High blood pressure Asthma Abnormal heart rhythm Heart attack diagnosis Diabetes Lung disease Arthritis Osteoporosis
Model III	Mobility_num IADL_num ADL_num ADL: Bathing/showering IADL: Using map in strange place Mobility: Stooping, or kneeling ADL: Getting in/out of bed ADL: Climbing one flight stair IADL: Managing money ADL: Dressing	Mobility_num IADL_num ADL_num Mobility: Sitting for 2 hours Mobility: Picking up 5p coin Mobility: Walking 100 yards Mobility: Climbing one flight stair ADL: Eating IADL: Using map in strange place ADL: Dressing
Model IV	Depression Closeness to spouse Marital status Age Mobility_num Money shortage Health problem Psychiatric condition Contact with children Contact with friends	Age Contact with friends Contact with children Closeness to spouse Money shortage Education Alcohol: how-often Health problem Self-reported hearing Mobility_num
Model V	Loneliness (prev.) Depression People No. in household Closeness to spouse Marital status Closeness to spouse (prev.) Marital status (prev.) Age (prev.) Age Contact with children	Age Closeness to spouse Contact with friends Money shortage Contact with children Age (prev.) Health problem Alcohol: how-often Loneliness (prev.) Contact with children (prev.)

Table 11. Performance comparison in different predictive models at the variable category level

AUC (validation/test)				
	Single-item measure of loneliness		**Multi-item measure of loneliness**	
	XGBoost	**LightGBM**	**XGBoost**	**LightGBM**
Model I	0.851/0.844	0.856/0.849	0.808/0.813	0.790/0.795
Model II	0.604/0.586	0.618/0.604	0.608/0.603	0.602/0.597
Model III	0.684/0.664	0.686/0.666	0.644/0.639	0.634/0.628
Model IV	0.853/0.848	0.859/0.851	0.812/0.816	0.798/0.804
Model V	0.878/0.882	0.879/0.886	0.854/0.874	0.844/0.859
Logloss (validation/test)				
	Single-item measure of loneliness		**Multi-item measure of loneliness**	
	XGBoost	**LightGBM**	**XGBoost**	**LightGBM**
Model I	0.262/0.251	0.260/0.249	0.400/0.363	0.409/0.371
Model II	0.351/0.337	0.347/0.335	0.494/0.455	0.495/0.456
Model III	0.335/0.323	0.334/0.323	0.485/0.447	0.489/0.451
Model IV	0.260/0.249	0.256/0.248	0.398/0.361	0.406/0.369
Model V	0.241/0.225	0.242/0.225	0.357/0.311	0.364/0.322

Table 12. Top-ranked variables in Model IV (full data in current wave) at the variable category level

	Single-item measure of loneliness	**Multi-item measure of loneliness**
	XGBoost Model	**XGBoost Model**
1	Depression[No]	Age
2	Marital status[Widowed]	Mobility_num
3	Closeness to spouse[N/A]	IADL_num
4	Depression[Yes]	Money shortage[Often]
5	Closeness to spouse[Very close]	Gender
6	Marital status[Married]	ADL_num
7	Age	Contact with friends[[1-2 time a week]
8	Mobility_num	Contact with friends[N/A]
9	Money shortage[Often]	Money shortage[Never]
10	Closeness to spouse[Quite close]	Contact with children[>=3 times a week
11	Psychiatric condition[N/A]	Health problem[Never]
12	Contact with friends[N/A]	Contact with children[1-2 time a week]
13	IADL_num	Depression[Yes]
14	ADL_num	Closeness to spouse[Very close]
15	Gender[Male]	Contact with friends[>=3 times a week]
16	Pulling/pushing large object[No]	Money shortage[Sometimes]
17	Alcohol[No]	Education[No qualification]
18	Contact with children[N/A]	Self-reported eyesight[Good]
19	Mild sport/activity[>1 times a week]	Health problem[Often]
20	Health problem[Never]	People No. in household[2-person]
	LightGBM Model	**LightGBM Model**
1	Age	Age
2	Mobility_num	Mobility_num
3	Money shortage[Often]	IADL_num
4	IADL_num	Gender
5	Gender[Male]	ADL_num
6	Contact with friends[N/A]	Money shortage[Often]
7	Marital status[Widowed]	Contact with friends[N/A]
8	ADL_num	Health problem[Never]
9	Closeness to spouse[Very close]	Contact with children[>=3 times a week]
10	People number in household[1-person]	Contact with children[1-2 time a week]
11	Psychiatric condition[Yes]	Contact with friends[1-2 time a week]
12	Psychiatric condition[N/A]	Marital status[Widowed]
13	Contact with children [N/A]	Money shortage[Never]
14	Education[No qualification]	Contact with friends[>=3 times a week]
15	Self-reported health[Fair]	Depression[Yes]
16	Self-reported health[Poor]	Psychiatric condition[N/A]
17	Closeness to spouse[Quite close]	Education[No qualification]
18	Vigorous sport/activity[Hardly/never]	Health problem[Often]
19	Climbing one flight stair[No]	Closeness to spouse[Close]
20	Health problem[Often]	Self-reported eyesight[Good]

Table 13. Top-ranked variables in Model V (full data in both previous and current wave) at the variable category level

	Single-item measure of loneliness	**Multi-item measure of loneliness**
	XGBoost Model	**XGBoost Model**
1	Loneliness[Yes] (prev.)	Age
2	Depression[No]	Age (prev.)
3	Depression[Yes]	Money shortage[Often]
4	People No. in household[1-person]	Closeness to spouse[Very close]
5	Closeness to spouse[Very close]	Depression[Yes]
6	Closeness to spouse[N/A](prev.)	Mobility_num
7	Closeness to spouse[N/A]	Mobility_num (prev.)
8	Marital status[Widowed]	Health problem[Often]
9	Age	Closeness to spouse[Not very close]
10	Mobility_num	Health problem[Never]
11	Money shortage[Often]	Gender
12	Marital status[Widowed] (prev.)	Loneliness[No] (prev.)
13	Age (prev.)	Loneliness[Yes] (prev.)
14	IADL_num	Contact with friends[N/A]
15	Contact with friends[N/A]	IADL_num
16	ADL_num	ADL_num
17	Gender[Male]	Contact with friends[>=3 times a week]
18	Walking 100 yards[No]	Depression[No]
19	Closeness to spouse[Not very close]	Contact with children[>=3 times a week]
20	Closeness to spouse[Very close] (prev.)	Contact with children[1-2 time a week]
	LightGBM Model	**LightGBM Model**
1	Age	Age
2	Mobility_num	Mobility_num
3	Mobility_num (prev.)	Money shortage[Often]
4	Closeness to spouse[Very close]	Depression[Yes]
5	Money shortage[Often]	Closeness to spouse[Very close]
6	Loneliness[Yes] (prev.)	Health problem[Never]
7	People number in household[1-person]	Health problem[Often]
8	ADL_num (prev.)	Loneliness[N/A] (prev.)
9	Age (prev.)	Mobility_num (prev.)
10	Marital status[Widowed]	IADL_num
11	IADL_num (prev.)	ADL_num
12	IADL_num	IADL_num (prev.)
13	Gender[Male]	Closeness to spouse[Close]
14	Closeness to spouse[N/A] (prev.)	Closeness to spouse[Not very close]
15	ADL_num	Gender
16	Depression[No]	Contact with friends[N/A]
17	Religion[No]	ADL_num (prev.)
18	Depression[Yes] (prev.)	Age (prev.)
19	Contact with children[N/A]	Contact with children[1-2 time a week]
20	People No. in household[2-person] (prev.)	Contact with friends[>=3 times a week]

Table 14. The correlation of prediction results between the XGBoost and LightGBM on the test data (wave-7) using Model V

	Single-item measure of loneliness	Multi-item measure of loneliness
AUC	0.9807	0.9091
Logloss	0.9555	0.8885

Table 15. The chi-squared test (χ^2) results between the outcomes using single-item and multi-item measures of loneliness

	Wave-2	Wave-3	Wave-4	Wave-5	Wave-6	Wave-7
χ^2 score	1600.6	1577.1	1728.0	1850.1	1709.3	1597.3
P-value	0.000	0.000	0.000	0.000	0.000	0.000

Note: The 95% Confidence Interval (CI) level is used.

Figures

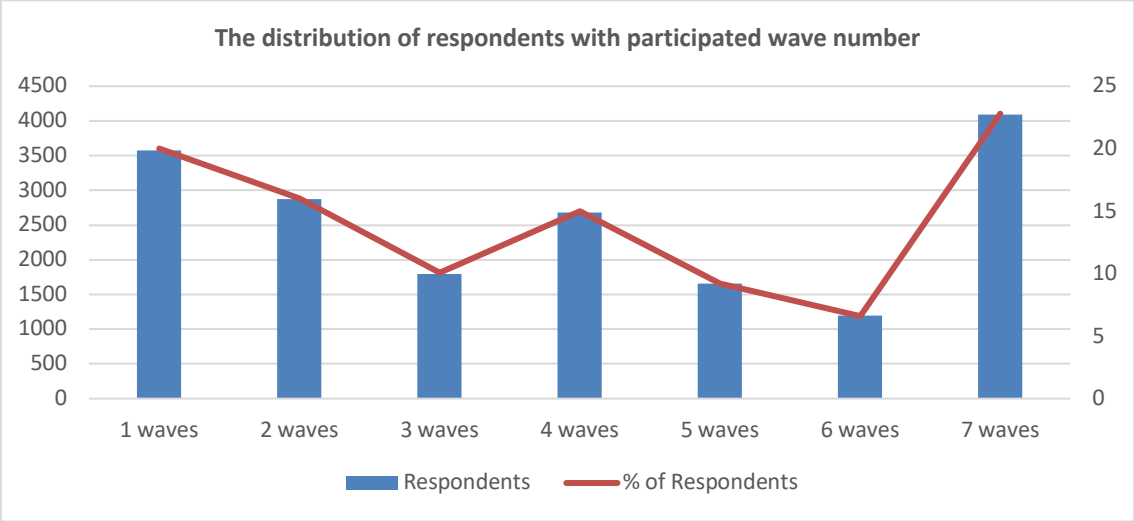

Figure 1. The distribution of respondents (N=17,861) according to the number of waves they participated in.

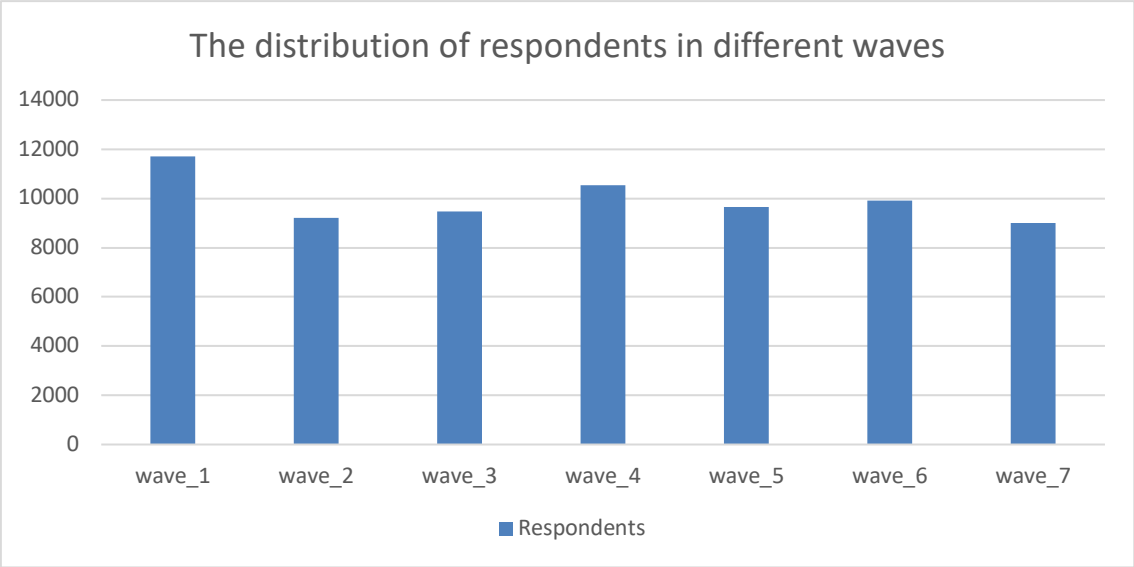

Figure 2. The distribution of respondents in different waves

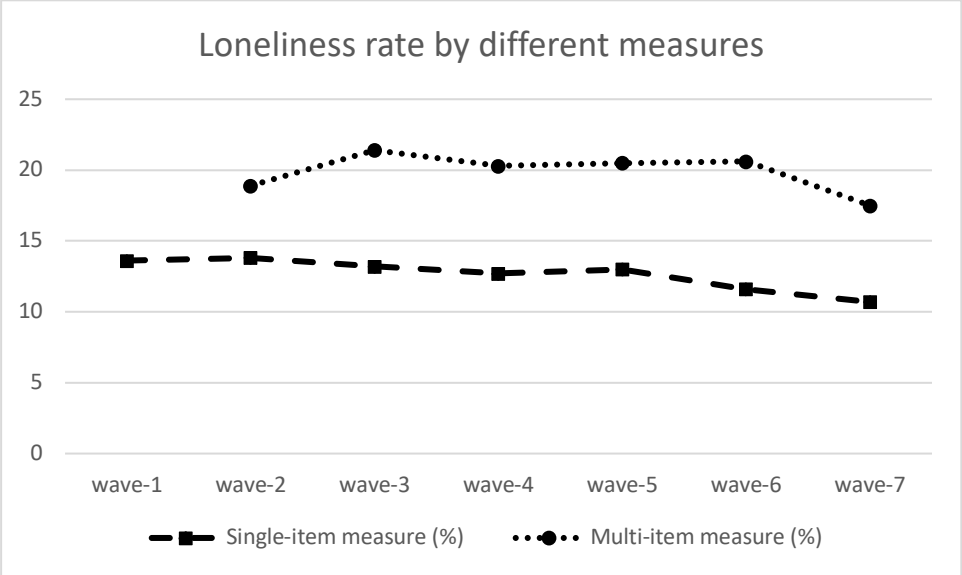

Figure 3. The comparison of loneliness prevalence rate in different waves

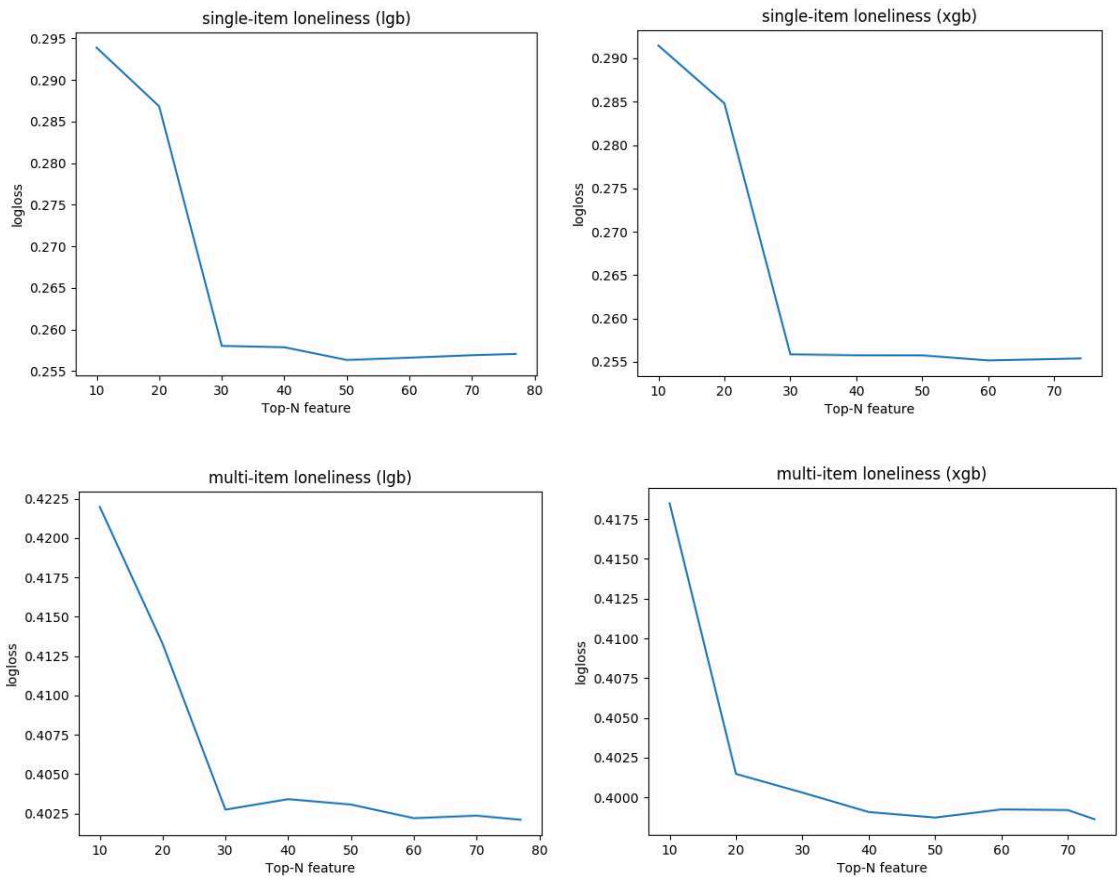

Figure 4. The relationship between the Logloss error and the top-ranking feature number in terms of different ML algorithms and loneliness measures