Wassim Ben Youssef
Student ID: 160043583

Research Project Proposal
Supervisor: Dr. Artur D'Avila Garcez

Prediction of Salary from job ads

Introduction

The aim of this project is to find a way to predict the salary of jobs in the UK based on job ads. The idea is to use descriptions provided in job ads (with features like "title", "skills" ...) to predict the salary. The job offer descriptions dataset was built from job ads found on the recruitment website reed.co.uk.

The products that will be generated are a predictive model of salary based on job offer description. The several stages of the research will be:

- Clean of the dataset containing features for job description and job salary
- Build the best predictive model by comparing different methods
- Analyzing the results and trying to improve them

The research question for this project is:

How can I predict the salary of online job ads? How can I improve the model to find best results?

The principal beneficiaries of this work would be:

- Bank and Insurance companies which can improve their current income prediction models. This can also help them to improve their credit risk checking of a customer based on salary prediction. The prediction can also help them for product recommendation based on incomes.
- Recruitment websites such as Glassdoor, Reed, Indeed. The model can help them to improve their recommendation algorithms. It can also be useful to figure out the market worth.
- Pensions which can use it for product recommendation and for risk evaluation.
- Employers and job seekers to figure out the market worth of different positions.

Critical context

The use of salary prediction

Few salary prediction models have been built to serve different purposes. The first to undertake salary prediction from job ads was Adzuna, a company that predicts the salary of an individual using the data contained in his CV. This application allows users to have an estimation of the salary they can expect when applying for jobs. Job prediction also helps employers to figure out and have an overview of the market worth [1]. In [2], salary prediction was also used to figure out the salary of graduate students. The objective of this study was to motivate students to work harder, knowing that a well-paid job may await them after they graduate. This can help recruitment website to improve the experience of users searching jobs, and help employers and job seekers to have a better understanding of the market worth.

Salary is also considered as an important factor to determine success in life [3]. It is then a relevant factor used in Bank, Insurance and Pension industry to calculate several types of risks, for instance: financial risk for insurances and credit risk for banks. The value of individuals salary also helps pension actuaries to establish pension plans for their customers [4]. It also helps these institutions to target which product should be proposed to specific individuals. However, it is difficult for these financial institutions to figure out the incomes of their customers due to the fact that these ones rarely share this information [4]. Hence, these financial institutions build models to predict incomes of their customers using the data they have on them, in the objective of optimizing their decision-making systems.

Previous works

To build their system, Adzuna has begun by launching a Kaggle competition in which participants had to build a regression model to predict job salaries. The dataset used was based on job ads in the UK taken from recruitment websites [5]. The 6 features of the dataset provide details on the jobs like "Full description", "Location" or "Title". Some participants have shared their code online on GitHub. All of them have used Random Forest algorithm to obtain their best accuracy.

P. Khongchai and P. Songmuang [6] have built a system to help students to predict their salary when they will graduate. The system was built using profiles of former students as training set, using several independent features like "Gender", "Faculty" and "Program". The predictive variable was a categorical variable with four levels as classes, each one being an interval of the salary. The authors then compared results with different methods and the best accuracy was found with Random Forest.

Liu et al. [7] have built a Bayesian Regression model to figure out the influence of factors like gender, race education, gifted or non-gifted student on yearly income in the US. The dataset used was built from a survey on 4 years where "subjects were nationally representatively sampled eighth-grade students".

Kibekbaev and Duman [8] have compared a mix of 16 linear and non-linear regression techniques on real-life dataset to predict bank customers incomes.

Unsupervised Learning

The data that can be found on job adds is very sparse and messy, especially when it is text data. We will then need to use unsupervised learning algorithms in order to build another representation of the data. Hence, unsupervised learning is used to find patterns in unstructured data by clustering it or by reducing the dimension [13].

Transformation of text data into numerical data

Most of the data available on job offers are text. Text data is quite complicated to handle in its initial form, due to the diversity of the vocabulary and mistakes and punctuations that can be irrelevant for the study [14]. We hence need to transform it into numerical data to be able to use it for analysis and predictions. The first step is to filter the text(s) from all punctuations and elements not desired. We first tokenize the text: we transform the whole text into tokens of strings. The tokens are generally words, they can also be smileys. Then, we filter the tokens by deleting the tokens not relevant: punctuations, prepositions and article words (the, in, on …), numbers and other kinds of irrelevant words [15]. Depending on the context, the filter can change, for instance, some might prefer to keep punctuations to identify smileys.

Then, we have to assign a unique identifier to each word, that is called indexing. However, instead of doing so for each unique word, we give the same index to words sharing a common meaning. For instance, "consult", will be what we call the stem for "consultant", "consultancy", "consulting"… This process is called stemming.

Finally, we calculate the term-document matrix, in which each element $a_{i,j}$ is the multiplication of the frequency of the term i in the document j with the inverse document frequency. The inverse document frequency of the term i is defined by the formula $idf_i = log\frac{N}{df_i}$ where N is the total number of documents in the corpus and df_i is the number of documents in which the term i appears.

LSA for dimension reduction with text data

Latent Semantic Analysis is a method used to reduce the dimension of the term-document matrix described just before, by reducing the number of terms. Using the Singular-value decomposition, it computes components that reflect the patterns of the data [16], and hence ignore the smaller, less important influences. Then, if a word did not appear in a document but yet have an influence because the significance of this word end up close, the LSA will highlight it.

K-means clustering

The objective of the K-means clustering algorithm is to divide a dataset of M points and N dimensions into K clusters, and so to find K clusters for the M data points [18]. To do so, the algorithm finds k data points, called centers, by minimizing the mean squared distance from each data point to its closest center [17]. The data points are then clustered around their nearest center.

Regression algorithms

Regression algorithms allow predicting continuous values like temperature, incomes, GDP… using other known features. Several machine learning algorithms have been conceived to serve this purpose.

Linear regression

The linear regression method consists on trying to find the best linear function of the form

$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \cdots$ that allows to predict the Y using the independent variables $(X_1, X_2 \ldots)$ [19]. The best function is the one that minimizes the error between the predicted values and the actual values. This method is very efficient if we have a linear relation between the independent features and the predictable variable.

Methods based on decision trees

Decision trees regression

A simple definition of the decision tree would be "it is an analysis diagram, which can help decision makers to decide which is the best option between different options, by projecting possible outcomes. The decision tree, gives the decision maker an overview of the multiple stages by that will follow each possible decision" [21]. The goal of a decision tree regression is to explain a continuous value Y from other n discrete or continuous values X_k (k Є [|1,n|]). To do so, let say that we have m individuals/observations. The principle is to partition individuals by creating groups as much homogeneous as possible from the point of view of the target value, taking into account a hierarchy of the predictive capacity of the variables considered. This hierarchy allows to visualize results in a tree and to build explanatory rules.

Random Forest algorithm

The principal issue of a decision tree is its strong dependence with the sample, which means that it does not generalize well. To tackle this issue, the random forest algorithm builds several independent decision trees, using randomly selected samples of variables and individuals [20]. All the trees built are then put together. In a regression problem, it then takes the average of the trees to get predictions from new data. The advantages of this algorithm are that it does not overfit, the computation time is reasonable and it can handle thousands of predictors and hence it can handle noise [22]. This algorithm also provides a ranking of variables importance.

Gradient Boosting algorithm

The gradient boosting algorithm has a different approach from that of random forest. The trees are connected like series in such way that when a tree is built, it tries to improve the results of the tree that was built previously. Each iteration tries to correct the error of the previous iteration by "fitting a simple parameterized function (base learner) to current "pseudo"-residuals by least squares at each iteration" [23]. At the end, we get a linear combination of all the trees, and the ones with the lowest error are overweight.

Artificial Neural Networks

Haykin [24] has defined Artificial Neural Networks as follow: "A neural network is a massively parallel distributed processor made up of simple processing units, which has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network from its environment through a learning process.
2. Interneuron connection strength, known as synaptic weights, are used to store the acquired knowledge."

The artificial neural network is organized in layers of neurons. Each neuron in the network receives an input signal, processes this signal and then sends an output signal. The first layer is the input layer. Each node receives a vector containing a feature predictor values and these nodes are connected to the nodes of the next layers with synaptic weighted links.

Approaches: Methods & Tools for Analysis and Evaluation

Literature survey

A first literature survey has been done using Google Scholar and the City, University of London Library to have a good overview of the previous works and the different methods that can be applied during this study. A second literature survey must be done to find other methods of text analysis and data wrangling and cleansing.

Description of job ads dataset

The original dataset obtained from job ads from the website Reed.co.uk contains the following features:

- “date posted”
- “Title”: The title of the job, which is the first information we have when looking at an offer. It is also the title that is shown when we make a research using a bar search.
- “Type”: permanent or temporary
- “Location”: generally, it is the town and the county
- “Region”
- “Sector”
- “Parent sector”: the sectors clustered in a more general sector like “Financial service” or “Customer service”
- “Salary interval”: if the salary is given annually or monthly. We only use annual interval, we convert the monthly one into annual.
- “Salary min”
- “Salary max”
- “Salary difference”: difference between salary max and salary min
- “Recruiter”: the company that posted the ad
- “Description”: a detailed description of the offer, provided by the company that posted the offer
- “Skills”: a list of skills required for the job
- “Salary mean”: the target value, which is the mean between the

Analysis and Cleansing of Dataset

The first work is to make an analysis of the dataset: plots of histograms, box-plots, have a look at the number of unique values for categorical features, plot of word clouds … Then we will have to clean the data, especially by making some text analysis on the features such as “Title”, “Description”, “Locations”, “Skills” … To do so, we can use K-means clustering to reduce the dimension/number of unique values by regrouping the ones that are very close. We will also transform the text data into text-frequency matrix that we will then reduce the dimension using Latent Semantic Analysis. The LSA will create components that we will use as features instead of the text data.

Implementation of Regression Model

After cleaning the data, we will build several models to predict the incomes. We will begin with a linear regression model. Then we will use Random Forest and find the best hyperparameters using a Grid Search. We will also try the Gradient Boosting algorithm with a Grid Search to fit the hyperparameters. If the computation of the Grid Search takes too much time, we will use a Random Search. We will also make an analysis of the importance of each feature for the prediction. Finally, we will use an Artificial Neural Network algorithm to try to improve our results.

Evaluation and Comparisons

The models will be trained on training set, and the performances will be evaluated using 3 measures: the r^2, the mean squared error and the mean absolute error. To evaluate the models and to compare them, we will compute these measures on a test set.

Analysis of results to improve them

After obtaining an overview of which method seems to lead to best results, we will make an analysis of the results to find ways to improve it. For instance, we might modify the hyperparameters and try values that have not been tested before. Moreover, we might try other methods of text analysis, for instance Latent Dirichlet Analysis. We might also modify the number of components created by the Latent Semantic Analysis. We will then try to obtain a better model. Finally, we will analyze the final model and features to see if we can identify some patterns (for instance, maybe the word "Head" is correlated with high incomes).

Complete report

The report will be started at the beginning of the project and will be completed at each step. We will begin when the first clean and analysis of the data will be done. We will then continue after having implemented the models and comparing them. The last part will be done after having obtained the final model. The six last weeks of the project will be only focused on the report. Several drafts will be sent to the supervisor to have feedbacks and comments.

Additional Time

If additional time remains at the end of the project, a research will be done on how we can use this model to predict incomes from Social Media data. The study will include a literature review and a review of previous works and related works.

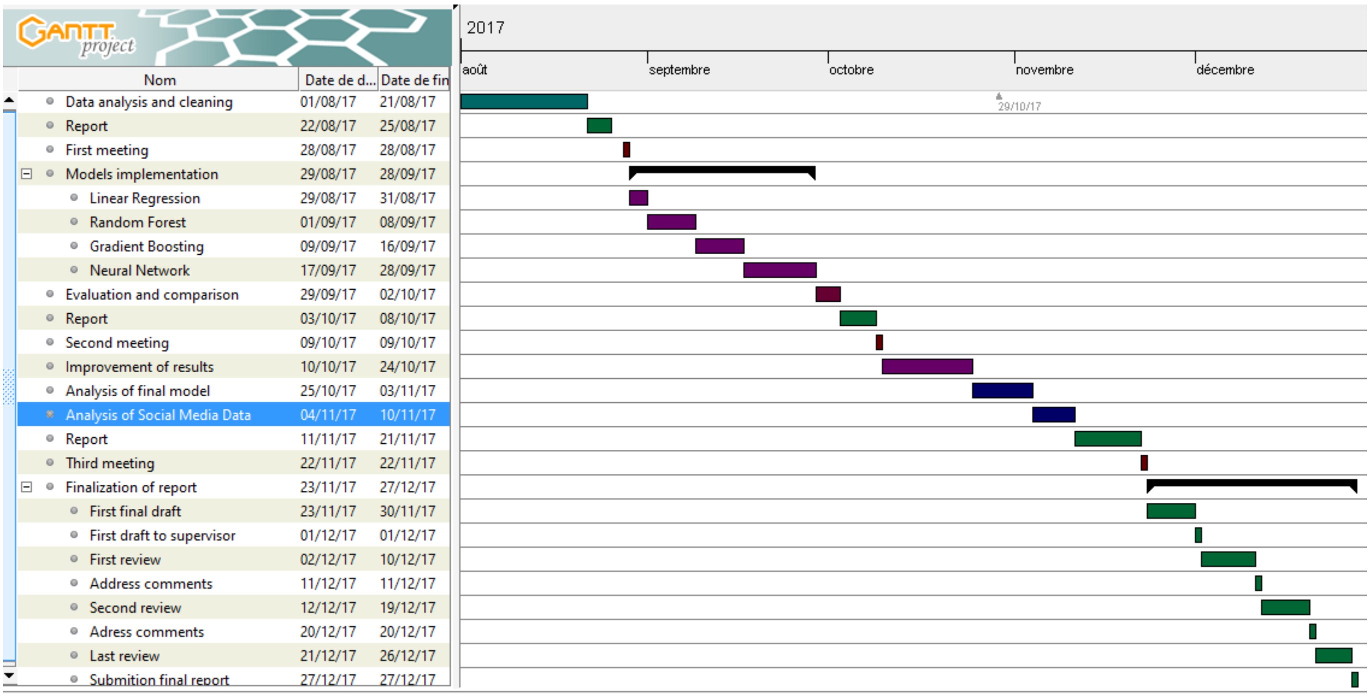
Project meetings

Several short meetings would be scheduled if issues are encountered and if needed. Three formal meetings will be scheduled at the end of each key step of the project to present formally what have been done.

Tools used

The study will be done using Python 3.6. We also will use Microsoft Azure to have more resources for computation, if needed.

Work Plan



Nom	Date de d...	Date de fin
Data analysis and cleaning	01/08/17	21/08/17
Report	22/08/17	25/08/17
First meeting	28/08/17	28/08/17
⊟ Models implementation	29/08/17	28/09/17
Linear Regression	29/08/17	31/08/17
Random Forest	01/09/17	08/09/17
Gradient Boosting	09/09/17	16/09/17
Neural Network	17/09/17	28/09/17
Evaluation and comparison	29/09/17	02/10/17
Report	03/10/17	08/10/17
Second meeting	09/10/17	09/10/17
Improvement of results	10/10/17	24/10/17
Analysis of final model	25/10/17	03/11/17
Analysis of Social Media Data	04/11/17	10/11/17
Report	11/11/17	21/11/17
Third meeting	22/11/17	22/11/17
⊟ Finalization of report	23/11/17	27/12/17
First final draft	23/11/17	30/11/17
First draft to supervisor	01/12/17	01/12/17
First review	02/12/17	10/12/17
Address comments	11/12/17	11/12/17
Second review	12/12/17	19/12/17
Adress comments	20/12/17	20/12/17
Last review	21/12/17	26/12/17
Submition final report	27/12/17	27/12/17

Risks

Description	Likelihood (0-5)	Consequence (0-5)	Impact (LxC)	Mitigation
Loss of data	1	5	5	Have a backup on GitHub
Some models take too much time to compute	4	3	12	Use a Virtual Machine on Microsoft Azure with high resources
Duration of some tasks is underestimated	2	4	8	Do not make an analysis on social media data if not enough time
Results are not well interpretable	3	2	6	Not a big issue since results are good enough. Try to explain as much as we can

References

[1] N. Yasmin and K. Kavinilavurajan, "Salary Prediction using Big Data", *International Journal for Scientific Research & Development* Vol. 4, Issue 01, 2016

[2] P. Khongchai and P. Songmuang, "Improving Students' Motivation to Study using

Salary Prediction System", *13th International Joint Conference on Computer Science and Software Engineering (JCSSE),* 2006

[3] S. Banerjee, "Why banks and other financial institutions predict income of their customers". Available: https://www.projectguru.in/publications/why-banks-and-other-financial-institutions-predict-income-of-their-customers/ [Accessed 01 July 2017]

[4] C. M. Bone and O. S. Mitchell, "Building Better Retirement Income Models" *North American Actuarial Journal*, *1*(1), 10–11, 1997

[5] Kaggle, https://www.kaggle.com/c/job-salary-prediction [Accessed 01 July 2017]

[6] P. Khongchai and P. Songmuang, "Random Forest for Salary Prediction System to

Improve Students' Motivation", *12th International Conference on Signal-Image Technology & Internet-Based Systems,* 2016

[7] H. Liu,Y-L. Kuo, Y. He and J. Liu, Yi, "Bayesian Regression Model For Predicting Income", 2006

[8] A. Kibekbaev and E. Duman, "Benchmark regression algorithms for income prediction modeling", Information Systems, 60 40-52, 2016

[9] H. Schoen, D. Gayo-Avello, P. T. Metaxas, E. Mustafaraj, M. Strohmaier, P. Gloor, "The power of prediction with social media", Internet Research, Vol. 23 Iss: 5, pp.528- 543. 10.1108/IntR-06-2013-0115, 2013

[10] H. He, A. Subramanian, S. Choi, P. K. Varshney and T. Damarla, "Social media data assisted inference with application to stock prediction", *Signals, Systems and Computers, 2015 49th Asilomar Conference on,* 10.1109/ACSSC.2015.7421462, 2016

[11] S. Asur, and B. A. Huberman, "Predicting the Future with Social Media", *Web Intelligence and Intelligent Agent Technology* (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, 2010

[12] K. Evangelos, T. Efthimios, T. Konstantinos, "Understanding the predictive power of social media", Internet Research, Vol. 23 Issue: 5, pp.544-559, 2013

[13] Z. Ghahramani, "Unsupervised learning." *Advanced lectures on machine learning*. Springer Berlin Heidelberg, 72-112, 2004

[14] F. Mahmoud and S. M. Benslimane. "Studying the effects of conflicting tokenization on LSA dimension reduction." *Multimedia Computing and Systems (ICMCS), 2014 International Conference on*. IEEE, 2014.

[15] Magerman, Tom, Bart Van Looy, and Xiaoyan Song. "Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications." *Scientometrics* 82, no. 2 (2009): 289-306.

[16] Deerwester, Scott, et al. "Indexing by latent semantic analysis." *Journal of the American society for information science* 41.6 (1990): 391.

[17] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. & Wu, A.Y. 2002, "An efficient k-means clustering algorithm: analysis and implementation", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 7, pp. 881-892.

[18] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979): 100-108.

[19] Weisberg, Sanford, 1947, and Wiley Online Library EBS. *Applied Linear Regression.* Wiley-Interscience, Hoboken, N.J, 2005.

[20] Breiman, Leo. "Random Forests." *Machine Learning*, vol. 45, no. 1, 2001, pp. 5-32.

[21] N. I. Khawaldah and N. A. Ishtayeh, "Binary Decision Tree", *Al Zaraqa University*

[22] Cutler, Adele. "Random Forest for Regression and Classification", Utah State University, 2010.

[23] Friedman, Jerome H. "Stochastic Gradient Boosting." *Computational Statistics and Data Analysis*, vol. 38, no. 4, 2002, pp. 367-378.

[24] Haykin, S. S. 1931- (Simon Saher). *Neural Networks: A Comprehensive Foundation.* Prentice Hall, Upper Saddle River, N.J, 1999.

Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

Part A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

Part B: Ethics Proportionate Review Form. Students who have answered "no" to questions 1 – 18 and "yes" to question 19 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in this case. The approval may be provisional: the student may need to seek additional approval from the supervisor as the project progresses.

A.1 If your answer to any of the following questions (1 – 3) is YES, you must apply to an appropriate external ethics committee for approval.		*Delete as appropriate*
1.	Does your project require approval from the National Research Ethics Service (NRES)? For example, because you are recruiting current NHS patients or staff? If you are unsure, please check at http://www.hra.nhs.uk/research-community/before-you-apply/determine-which-review-body-approvals-are-required/.	**No**
2.	Does your project involve participants who are covered by the Mental Capacity Act? If so, you will need approval from an external ethics committee such as NRES or the Social Care Research Ethics Committee http://www.scie.org.uk/research/ethics-committee/.	**No**
3.	Does your project involve participants who are currently under the auspices of the Criminal Justice System? For example, but not limited to, people on remand, prisoners and those on probation? If so, you will need approval from the ethics approval system of the National Offender Management Service.	**No**

A.2 If your answer to any of the following questions (4 – 11) is YES, you must apply to the City University Senate Research Ethics Committee (SREC) for approval (unless you are applying to an external ethics committee).		*Delete as appropriate*
4.	Does your project involve participants who are unable to give informed consent? For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf?	**No**
5.	Is there a risk that your project might lead to disclosures from participants concerning their involvement in illegal activities?	**No**
6.	Is there a risk that obscene and or illegal material may need to be accessed for your project (including online content and other material)?	**No**

7.	Does your project involve participants disclosing information about sensitive subjects? For example, but not limited to, health status, sexual behaviour, political behaviour, domestic violence.	**No**
8.	Does your project involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning? (See http://www.fco.gov.uk/en/)	**No**
9.	Does your project involve physically invasive or intrusive procedures? For example, these may include, but are not limited to, electrical stimulation, heat, cold or bruising.	**No**
10.	Does your project involve animals?	**No**
11.	Does your project involve the administration of drugs, placebos or other substances to study participants?	**No**

A.3 If your answer to any of the following questions (12 – 18) is YES, you must submit a full application to the Computer Science Research Ethics Committee (CSREC) for approval (unless you are applying to an external ethics committee or the Senate Research Ethics Committee). Your application may be referred to the Senate Research Ethics Committee.		*Delete as appropriate*
12.	Does your project involve participants who are under the age of 18?	**No**
13.	Does your project involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.	**No**
14.	Does your project involve participants who are recruited because they are staff or students of City University London? For example, students studying on a specific course or module. (If yes, approval is also required from the Head of Department or Programme Director.)	**No**
15.	Does your project involve intentional deception of participants?	**No**
16.	Does your project involve participants taking part without their informed consent?	**No**
17.	Does your project pose a risk to participants or other individuals greater than that in normal working life?	**No**
18.	Does your project pose a risk to you, the researcher, greater than that in normal working life?	**No**

A.4 If your answer to the following question (19) is YES and your answer to all questions 1 – 18 is NO, you must complete part B of this form.		
19.	Does your project involve human participants or their identifiable personal data? For example, as interviewees, respondents to a survey or participants in testing.	**No**