Machine learning classification of entrepreneurs in British historical census data

Montebruno, Piero and Bennett, Robert and Smith, Harry and van Lieshout, Carry

Department of Geography, University of Cambridge, Cambridge Group for the History of Population and Social Structure (Campop)

2 August 2019

Machine learning classification of entrepreneurs in British historical census data

Piero Montebruno[a], Robert J. Bennett[a,*], Harry Smith[a], Carry van Lieshout[a]

[a] *University of Cambridge, Department of Geography and Cambridge Group for the History of Population and Social Structure, Downing Place, Cambridge, CB2 3EN, UK.*

Abstract

Thanks to recent data availability, digitized transcriptions of Victorian censuses provide unprecedented historical big data on individuals in the past, but also with new methodological challenges like the classification of otherwise underreported entrepreneurs among a population sample of millions of individuals. This paper presents a methodological solution to accomplish the task of classifying entrepreneurs. We apply machine learning, including deep learning, to outperform a standard logistic regression algorithm. Our methodological developments traverse traditional disciplinary lines using state-of-the-art artificial intelligence methods. The main conclusion of the paper is that significant gains in performance can be achieved with historical archive data through machine learning to test economic theories on historical entrepreneurship. This suggests applicability to other disciplines in information sciences.

Keywords: machine learning, deep learning, logistic regression, classification, big data, census

⋆

Declarations of interest: none

*Corresponding author: Tel.: +44-01223-339957; fax: +44-01223-333392;

Email addresses: pfm27@cam.ac.uk (Piero Montebruno), rjb7@cam.ac.uk (Robert J. Bennett), hjs57@cam.ac.uk (Harry Smith), cv313@cam.ac.uk (Carry van Lieshout)

1. Introduction

Modern information processing techniques are as applicable to classifying and identifying patterns in historical data as they are to modern data. An important historical question has been 'who were entrepreneurs in the past?' This is a first and essential step towards then identifying their characteristics and understanding their behaviour. The analysis of historical developments in entrepreneurship has lacked until recently sufficient data to be confident about the scale of historical activity and trends over time. After a major efforts of transcription and data coding large scale historical sources are now becoming available that allow entrepreneurs to be identified in the past from their descriptions of themselves. In England and Wales, a digitized version of the Victorian censuses over 1851-1911 has become available through the I-CeM database (Higgs and Schürer, 2014; Schürer et al., 2015). This has been enhanced in a supplementary database (the British Business Census of Entrepreneurs, BBCE) that extracts the members of the population who can be identified as entrepreneurs (Bennett et al., 2019c). This provides a new resource for information analysis, and also introduces scope to make long-term comparisons between modern and previous historical patterns. Unfortunately for the first four of these censuses (1851-81), accounting for nearly 80 million people, only a limited question referring to employers was used by the census administrators which does not allow direct and full identification of all entrepreneurs.

This paper studies the methodological challenge of classifying the employment status from the information that was self-reported in the census. Classification methods are assessed that are based on the individuals' demographics and also the descriptive text of their occupational activities in the archival records of the Census Enumerators Books (CEBs). Many learning methods have been developed in information science for such text-based classification; e.g. binary linkage (Boutell et al., 2004), classifier chains (Read et al., 2011), label powerset (Tsoumakas et al., 2011), rankings by pairwise comparison (Hüllermeier et al., 2008; Fürnkranz et al., 2008). These developments have expanded the focus in textual processing from title searches and tagging (Hu et al., 2006) to multiple tag interactions (Murthy and Gross, 2017; Al-Salemi et al., 2019; Tang et al., 2019), complex text interlinkages for result caching (Kucukyilmaz et al., 2017), deep textual semantic interactions (Kastrati et al., 2019), and attempts to identify sentiments through textual recurrence (Abdi et al., 2019).

The descriptive text used by entrepreneurs in the census that were explicitly identified in the census can be used to train machine learning methods to identify others in the census who were not explicitly identified or did not fully respond to the census questions. This classification of the population is of importance for understanding the scale and trends of entrepreneurship. The economic theory of entrepreneurship relies on the classification of individuals as Employer (E) and Own-account (OA) which was the Victorian census term for those proprietors operating on their own with no employees, as distinct from Workers (W). The sum of E and OA gives all self-employed, which following Parker (2004) and Blanchflower and Oswald (1998) we use as the definition of all Entrepreneurs (Ents). The methodology developed here for entrepreneurs is focused on evaluating alternative estimation methods for this classification. However, the paper has broader relevance for any classification process attempted in other disciplines. Thus, it is not restricted to economic history or historical data. The paper uses new developments in artificial intelligence (AI). AI refers to computers' thinking as humans do; as defined by The Editors of the American Heritage Dictionaries (2011), the verb to think can be defined as: "To exercise the power of reason, as by conceiving ideas, drawing inferences, and using judgment". This can be expanded by using training that involves known patterns as the input of the learning process where the patterns do not have explicit rules when computerised this is called a form of machine learning (ML). The process can be further expanded to deep learning (DL) when the model used is a distillation over several layers, or filters, where each attempts a better representation of the data often called a neural network because its inspiration comes from understanding of how the brain learns, though neural networks are not themselves considered a representation of the brain. François Chollet 2018 provides the following useful relation:

$$\text{Artificial Intelligence} \supset \text{Machine Learning} \supset \text{Deep Learning}$$

The methodology developed in this paper uses AI, ML, and DL to tackle a problem that derives from the way in which census administrators collected information in the nineteenth century. The historical mid-Victorian censuses in the UK that are now available as a digital database were collected in two formats: first, for 1851-81 a question was used that sought to distinguish employers and 'masters' from others: i.e. those individuals who were able to operate on own account either as sole operators, or employing others; second, for 1891-1911 the question was modified to ask individuals explicitly to identify themselves as employers, own account, or workers: termed their 'employment status'. Hence, for the later period the question attempted to collect full information on all entrepreneurs (as employers or own account) for the whole population. The change in the questions was a response to pressures from social scientists led by Charles Booth and Alfred Marshall, that the census administrators (General Register Office: GRO) introduce a new question that identified the self-employed (Treasury Committee, 1890; see also Higgs, 2004).

The result of this change was a major improvement in census design as it provided a separate classifier that explicitly identified entrepreneurs, which was additional to their textual description of their occupation. However, it created a discontinuity with the earlier period where the census question provides potentially full coverage of employers, but only partial coverage of own account for those cases where they identified themselves as 'masters'. The term 'master' had an historical meaning for those trained or apprenticed in some trades who could operate alone or employ others, but it was a term that was obsolete in many occupations by the mid-Victorian period, whilst in other occupations 'master' had never been used (for example in professions, commerce, transport, and many retail trades). Indeed, in 1851 only about 6 per cent of entrepreneurs used the term master, which fell to about 3-4 per cent by 1881 (Bennett et al., 2018, 2019b). The classification problem that we tackle is: can the information in the later censuses on 'employment status' be used to train an information classifier to identify entrepreneurs in the early censuses using the textual responses to the question on employers, masters, and other occupations? Also, can information gathered from a subsample of the early censuses be used to train a classifier to generalize and identify entrepreneurs in the early censuses using standard demographic features, or alternatively the very detailed occupational information strings that individuals used to describe themselves? Finding a way to estimate entrepreneurial status for this early period is an important challenge since the later census questions align closely with modern censuses, thus allowing a continuous series of to be developed from 1891 to the present. Having an available benchmark for entrepreneurial status for 1851-81 allows the time series to be extended from 1851 up to the present; and it would also help develop long-term comparisons backwards to earlier periods before 1851.

Despite a progressive adoption of machine learning, this paper is one of the first to apply machine learning in an historical setting. Moreover, this use of machine learning solves a methodological gap in the classification of millions of individual that on the night of each census responded with valuable demographic and economic information. In this paper we describe the classification problem, present the methodology for applying ML, and test the performance of different ML algorithms.

2. Methodology

The machine learning method we develop seeks to tackle a binary classification problem (if the labels are W and Ent), or a multi-class classification problem (if the labels are W, E and OA) (Boutell et al., 2004; Tsoumakas et al., 2011; Read et al., 2011). We test the performance of different ML algorithms against a traditional probability based model using Logistic Regression (LR). The LR has been used in many information processing applications and has been the algorithm of election previously used to tackle the problem at hand by (Bennett et al., 2018, 2019b). It is used here as a benchmark for comparison. The LR method has the advantage of using characteristics of the individuals themselves to classify their entrepreneurial status: for example, their age, gender or household status in relation to other individuals in the same household (e.g. a daughter, son, aunt, or someone lodging). This is a powerful method, but it has two defects: first, it assumes that the individual characteristics that indicate entrepreneurial status remained similar over time; second, it ignores potentially rich information that is available in the occupational descriptors that individuals gave. For 1851-81 the full occupation string descriptors have a maximum of 300 characters that summarise people's economic activities. These text strings provide a level of deep content on firm sizes (such as number of employees) that can add considerable information to improve accuracy of classification over the LR method. Our method attempts to use these strings as an ML classifier applied to the digitized CEBs in order to separate individuals into employment statuses, giving both binary and multi-class classification for the earlier censuses.

There is a rapidly growing literature on machine learning in the information sciences. However, there have been few applications to economic history. Schürer et al. (2015) develop a computerized classification method for the same 1851-1911 I-CeM data as used in this paper: to standardize and code occupational titles and also birthplace descriptors. This uses dictionaries of occupations and birthplaces and then develops a hierarchical system of matching to link actual terminology to dictionary terms. However, this is an AI and not a ML method, though it tackles a similar problem to that here. Other applications of ML to related social science questions have used standard information science techniques such as Bayes Networks (Tang et al., 2016), used by Alvarez-Galvez (2016), to tackle interrelationships between socioeconomic status and health in Europe, or other machine learning methods used by Su and Meng (2016) to perform automated text analysis of online forums to assess the response to China's government policies, generalized boosting used by

Reichenberg and Berglund (2019) to overcome some of the deficiencies of an inverse-probability weighting analysis, and structured learning used by Katz and Levin (2018) to classify individuals into types of political supporters using ML, based on joint responses to eight questions while estimating the association between each item and support dimension.

The area of ML that we develop can be understood as *predictive* or *supervised* process of learning a mapping from inputs $\mathbf{x}$ to outputs y given a labeled set of inputs pairs $D = \{(\mathbf{x_n}, y_n)\}_{n=1}^{N}$ whith D the *training set*, and N the number of training examples (Murphy, 2012). In ML the inputs $\mathbf{x}$ are *features* (or *attributes*) while the outputs y are *labels* (or *targets*). When y is *nominal* or *unordered-categorical* with j categories and j goes from 1 to C, the problem is *classification* or *pattern recognition* (Murphy, 2012). If $C = 2$, the classification is *binary* and y is taken to be $\{0, 1\}$. If $C > 2$, the classification is *multi-class* (Murphy, 2012) . Traditional ML follows a method called *function approximation* where it is assumed that $y = f(x)$ for some unknown function f and the learning process is aimed at estimating the function f given a labeled training set, and then to make the predictions as follows:

$$\hat{y} = \hat{f}(\mathbf{x})$$

The process of calculating out-of-training-set predictions is then called *generalization*. Additionally, an algorithm that puts into action classification is called a *classifier*. It can also refer to the mathematical function performed by a classification algorithm, that maps features to labels. In our case, the chosen base-line classifier is the logit model (LR) for the binary responses and the multinomial logit (MNL) for the multi-class responses. This is a traditional classifier approach (e.g. Cheng and Hüllermeier (2009)). The logit models are applied by using the attributes of each individual that most closely correlate with the entrepreneurial status where this was recorded in the later censuses. This logit estimate is then applied to the earlier censuses to classify individuals where only partial records of entrepreneur status were recorded. This gives the binary probability of being an entrepreneur or not between Ent and W, or the multi-attribute probability of being E, OA or W.

As first established by Goldberger (1991) in his seminal book, the function of interest is the *conditional expectation function* (CEF) which in the case of a given value i of a binary label, y_i, is the probability that the label is 1 given the value of a feature i, x_i as presented in Rabe-Hesketh and Skrondal (2012):

$$E(y_i|\mathbf{x}) = \Pr(y_i = 1|\mathbf{x})$$

The probability must lie between 0 and 1, thus a non-linear link function is used to estimate the following linear relation (Rabe-Hesketh and Skrondal, 2012):

$$\text{link}\{\Pr(y_i = 1|\mathbf{x})\} = bias + weights' \, \mathbf{x}$$

where the intercept is called *bias* and the slope coefficients are called *weights* (Murphy, 2012). Sometimes bias and weights together are also called *parameters* (Goodfellow et al., 2016). The link function that we use in this paper is the *logit* defined as the logarithm of the *odds* by Rabe-Hesketh and Skrondal (2012):

$$\text{logit}\{\Pr(y_i = 1|\mathbf{x})\} \equiv \text{logarithm}\,\{odds(y_i = 1|\mathbf{x})\} = bias + weights' \, \mathbf{x}$$

and the odds that the label is one are defined as follows (Rabe-Hesketh and Skrondal, 2012):

$$odds(y_i = 1|\mathbf{x}) \equiv \frac{\Pr(y_i = 1|\mathbf{x})}{1 - \Pr(y_i = 1|\mathbf{x})}$$

Taking the inverse of the logit function makes possible to estimate the probability that the label is one given a certain value of the feature (Rabe-Hesketh and Skrondal, 2012):

$$\Pr(y_i = 1|\mathbf{x}) = \text{logit}^{-1}(bias + weights'\ \mathbf{x}) \equiv \frac{\exp^{bias+weights'\ \mathbf{x}}}{1 + \exp^{bias+weights'\ \mathbf{x}}}$$

Being $\Pr(y_i = 0|\mathbf{x}) = 1 - \Pr(y_i = 1|\mathbf{x})$, then also:

$$\Pr(y_i = 0|\mathbf{x}) = \frac{1}{1 + \exp^{bias+weights'\ \mathbf{x}}}$$

The logit model forms part of the so-called *single-index models* where the CEF is equal to a non-linear mean function $F()$ (i.e., the inverse of the logit, or logit^{-1}) of a single index, $weights'\ \mathbf{x}$, of the features and the weights, following Cameron and Trivedi (2005):

$$E(y|\mathbf{x}) = F(weights'\ \mathbf{x})$$

And the effect on the CEF of a change in the ith regressor is, according to Cameron and Trivedi (2005):

$$\frac{\partial E(y|x_i)}{\partial x_i} = F'(weights'\ \mathbf{x})weight_i$$

and $F' = \frac{\partial F()}{\partial}$. Thus the *relative effect* of changes in regressors is equal to the ratio of the weights, also following Cameron and Trivedi (2005):

$$\frac{\partial E(y|x_i)/\partial x_i}{\partial E(y|x_k)/\partial x_k} = \frac{weight_i}{weight_k}$$

as $F'(weights'\,\mathbf{x})$ cancels in the numerator and the denominator. This means that, for instance, if $weight_i$ is three times $weight_k$, then a one unit change of x_i has three times the effect of a one-unit change in x_k. And if $F()$ is monotonic, as the inverse logit is, then the signs of the weights command the signs of the effects for all possible values of the feature.

In the case of a multi-class label, for example, if we are classifying individuals into entrepreneurial status of E, OA or non-entrepreneurs as W, a general model can be written following Cameron and Trivedi (2005) where J is a given category out of the j categories and j goes from 1 to C:

$$\Pr(y_j = J|\mathbf{x}) = \frac{\exp^{bias^{[J]}+weights^{[J]}\,\mathbf{x}}}{\sum_{j=1}^{C}\exp^{bias^{[j]}+weights^{[j]}\,\mathbf{x}}}$$

where the superscript enclosed in the square brackets, [], is used to signal that the weights and the bias pertain to a given class, j and the denominator is equal to the sum of the numerators, so that the probabilities sum up to one. Consequently, this is a MNL with alternative-invariant features with alternative-specific weights. Also, the first, usually the most frequent category, is taken as base category which means the following two important assumptions: $bias^{[1]} = 0$ and $weight_i^{[1]} = 0$, so each weight must be interpreted as the change in probability that a unit increase in a given feature *with respect to* the base category. For the case, where $C = 3$ with categories $(j = 1) \equiv Worker$, $(j = 2) \equiv$ *Employer*, and $(j = 3) \equiv$ *Own account* the probabilities are as follows using Rabe-Hesketh and Skrondal (2012):

$$\Pr(y_j = 1|\mathbf{x}) = \frac{1}{1 + \exp^{bias^{[2]}+weights^{[2]}\,\mathbf{x}} + \exp^{bias^{[3]}+weights^{[3]}\,\mathbf{x}}}$$

$$\Pr(y_j = 2|\mathbf{x}) = \frac{\exp^{bias^{[2]}+weights^{[2]}\ \mathbf{x}}}{1+\exp^{bias^{[2]}+weights^{[2]}\ \mathbf{x}}+\exp^{bias^{[3]}+weights^{[3]}\ \mathbf{x}}}$$

$$\Pr(y_j = 3|\mathbf{x}) = \frac{\exp^{bias^{[3]}+weights^{[3]}\ \mathbf{x}}}{1+\exp^{bias^{[2]}+weights^{[2]}\ \mathbf{x}}+\exp^{bias^{[3]}+weights^{[3]}\ \mathbf{x}}}$$

This is a discrete choice model to predict the employment status of any economically active individual in the census using, as previously said, alternative-invariant features with alternative-specific weights.

3. Data

The data that we seek to classify are the transcriptions of the 1851-1911 censuses of England and Wales as provided in Higgs and Schürer (2014) supplemented as in Bennett et al. (2019c). For our purpose, we use only the non-farm population of entrepreneurs because the census data collected on farmers is sufficient to allow their entrepreneurial status to be identified without the need for machine learning or supplementation of census responses; hence they do not need the estimation processes discussed here (for a discussion of extracting farm entrepreneurs see van Lieshout et al. (2019) and for an assessment of shifts in agrarian entrepreneurs see Montebruno et al. (2019a)). For non-farm entrepreneurs, Table 1 confirms that, since the means and the medians for each feature are statistically significantly different for each entrepreneur label class for the later censuses (in this case for 1891 as an example), a binary classification can be used based on the features of 1891 as a training set. In fact, all the t- and z-statistics of the two-sample t-test with equal variances and the the two-sample Wilcoxon rank-sum, or Mann-Whitney, tests show that the difference between the means and the medians for the group are statistically significantly different from zero with p-values roughly equal to zero. At the same time, Table 2 shows a similar picture for classifying multiple attributes using the MNL with same features but now with labels 1 = Worker (W), 2 = Employer (E), and 3 = Own account (OA). Again the t- and z-statistics are all statistically significantly different form zero with corresponding p-values (almost) equal to zero.

4. Empirical analysis

[For referees only: The empirical analysis uses three ground-truth (gold standard) datasets that are attached to this submission as supplementary material only for the Reviewers to download; after acceptance of the manuscript, the data will be identified by IDs in a Mendeley Data deposit linked to the paper so that they can be downloaded from Higgs and Schürer (2014) and Bennett et al. (2019c). Redeposit of the datasets is not possible because of licensing restrictions but readers will be able to download the full datasets and replicate the results in this paper from the sources cited.] Our approach to the problem of classification of the 1851-81 censuses is to train the data with the known labels in the later 1891-1911 censuses, using the entrepreneur status that is fully reported in these later censuses (but not the earlier ones). Following the definition of predictive or supervised ML, we first develop a base-line model by approximating the unknown classification function with a logit (LR) model, using as training set the 1891 census where the labels 0 = Worker and 1 = Entrepreneur come from the reported employment status responses given in this census that are not available in earlier censuses. The LR classifier uses as classification features the following: the coding of the individual's occupational statement (SubOccode: See Bennett et al. (2018), for the list of the 844 occupational categories), Registration- sub-district (RSD) population Density (to use information of each individual's location), the individual's Age, Sex, Marital status, Relationship to the head of the household, and Number of servants in the household (which is a family resource surrogate). The method tests the accuracy of the LR for classifying individuals for whom their entrepreneur status is known from their 1891 census responses. The dataset has 1,000 entrepreneurs and 1,000 non-entrepreneurs (the training set is a random subset of 60% and the test set a random subset of the other 40% of the full datasets). The LR achieves an accuracy of 0.74 and a confusion matrix (see below for a full explanation) given in Figure 1a. The full trained model is presented in Table 3 where a logit classifier is used and each weight ($\boldsymbol{w}$) and its $\boldsymbol{t}$-statistic are given. The same model but stripped of the SubOccode feature for computing efficiency is given in Table 4 where not only weights but also partial derivative marginal effects are provided. Note that the marginal effects are given only for the level variables and not for the squared, interaction and constant terms.

A similar procedure is followed with a MNL model with labels W/E/OA dropping the SubOccode feature for ease of computation with a dataset of 1,000 in each category (and similarly defined training set as above). Table 5 shows each weight ($\boldsymbol{w}$) and its $\boldsymbol{t}$-statistic, while Table 6 shows the partial derivative marginal effects for each variable in levels.

The performance of the method can be improved by keeping the standard LR classifier but using a dataset for training and testing purposes from the 1851-81 censuses with labels derived not from full information of entrepreneurial status as in the later censuses where information of employment status was given in a second question but, instead, from clerical labeling by researchers of employment status using the occupation text strings of a large subsample of individuals as W/E/OA. The occupational strings are contained in the Higgs and Schürer (2014) dataset, with the clerical coding of entrepreneur status derived from Bennett et al. (2019c). The strings have terms such as House servant, "Piano forte maker master", "Boot and shoe maker employing 4 men", "Iron mine proprietor employing 70 men & engineer empg [sic] 100 men total 170 men" or "chimney sweeper empling [sic] 1 man & 213 boys". Here and below we use a sample from the 1851 census data containing individuals of all types; the dataset has 1,000 workers, and 500 each of employers and own account. The use of this 1851 dataset to train the LR method increases performance to 0.82 for the accuracy of estimating entrepreneurial status, as shown in the confusion matrix Figure 1b. The quest to keep improving on these methods is the main aim of the rest of the paper. As (Wolpert, 1996) established there is no universally best model (i.e. the no free lunch theorem), and the assumptions that work well in one problem do not necessarily work well in another. Thus our aim is to actively look for better performance among the inherent uncertainties of ML algorithms applied to this new research problem.

4.1. Classifier Comparison

Using Python library scikit-learn (Pedregosa et al., 2011), we compare a range of alternative classifiers to the standard logistic regression based on a code by Varoquaux and Müller (2018) (under a 3-clause BSD License). Figure 3 shows the accuracy and 2-D predicted probability grid for the label being an entrepreneur (Ent) with 2,000 balanced random data points for ten classification algorithms: Nearest Neighbors, Linear Support Vector Machine (SVM), Radial Basis Function (RBF) SVM, Gaussian Process, Decision Tree, Random Forest, Natural Network (Net), Adaptive Boosting (AdaBoost), Naive Bayes, and Quadratic Discriminant Analysis (QDA)) (see Zhang and

Zhou (2007); Schapire and Singer (1999); Tang et al. (2016); Tong and Chang (2001); Wu et al. (2014); Alvarez-Galvez (2016); Freund and Schapire (1996); Friedman (2001); Murphy (2012)). In the first row of the figure the features are Age and SubOccode, in the second row Density of the RSD and SubOccode, and in the third Density of the RSD and Age. The purple circles are Ws and the green ones Ents. The color in the background is the 2-D predicted probability grid which means that when the grid is purple a test point will be classified as W and when the color is green a test point will be classified as Ent. The resulting probability patterns are strikingly similar to known patterns from the data (seeBennett et al. (2019a)); e.g. that younger and older people are less entrepreneurial, and that lower density locations and certain SubOccodes are more entrepreneurial. The results show that the best performing methods are AdaBoost (which achieves accuracy of 0.94, 0.90, and 0.72, respectively, for all three rows), Decision Tree and Random Forest (both respectively 0.93, 0.91, and 0.72), and Gaussian Process (0.93, 0.89, 0.73). The standard LR performs systematically worse than almost all of the other methods tested here. Also, it can be seen that the best predictions are made using the features Age and SubOccode in combination, while the poorest predictions are made from Density of the RSD with Age. Of course, these are just three features for ease of visualization but our final model selection uses all available features in the dataset.

4.2. Confusion matrix

The assessment of the classification of Ents and Ws by Age and Suboccode using the benchmark LR can be visualized in Figure 2. There are two areas separated by a linear hyperplane. The first has a green zone on top and to the right where the probability indicates a likelihood of being an Ent with two sets of individuals: green circles or True Positives (TP), true Ents predicted as Ents, and light purple crosses or False Positives (FP), true Ws predicted as Ents. The second area has a purple zone at the bottom and to the left where the probability indicates a likelihood of being a W with again two sets of individuals: light green crosses or False Negatives (FN), true Ents predicted as Ws, and purple circles or True Negatives (TNs), true Ws predicted as Ws. Once we have selected the classifier, we run the AdaBoost method of Freund and Schapire (1996) using a parametrization suggested in Dawe (2018) (under a 3-clause BSD License) for the dataset with the maximum possible set of "extracted" (those labeled according to their strings or type of employment occupation codes) Ents (after excluding farmers). The method is now applied to the full 1851 labeled subsample of

70,872 Ws, 35,436 Es, and 35,436 OAs. In the binary classification problem, the following table shows at the bottom and in bold the predicted labels $(\widehat{-}, \widehat{+})$, on the left and in bold the actual labels $(-, +)$, four cells with TN, FP, FN, and TP:

	$\boldsymbol{N_{\widehat{-}}}$	$\boldsymbol{N_{\widehat{+}}}$	**TOTALS**
$-$	TN	FP	$\boldsymbol{N_-}$
$+$	FN	TP	$\boldsymbol{N_+}$
	$\widehat{-}$	$\widehat{+}$	

The confusion matrix is similar with the only difference that the number in the cells are rates over the previous table numbers after summation by rows. In particular it is important the sum of the rows or the true number of negatives, upper row or $\boldsymbol{N_- = TN + FP}$, and the true number of positives, lower row or $\boldsymbol{N_+ = FN + TP}$:

$-$	Specificity	False Alarm	(Denominator $\boldsymbol{N_-}$)
$+$	Missed Detection	Sensitivity	(Denominator $\boldsymbol{N_+}$)
	$\widehat{-}$	$\widehat{+}$	

Upper left, the Specificity Rate, $\boldsymbol{TN/N_-}$; upper right the False Alarm, Type I errors, or False Positive Rate, $\boldsymbol{FP/N_-}$; lower right, the Sensitivity, Recall, True Positive or Hit Rate, $\boldsymbol{TP/N_+}$; and lower left, the Missed Detection, Type II errors, or False Negative Rate, $\boldsymbol{FN/N_+}$. Not shown are the rates summing the columns for the "called" number of positives, right column or $\boldsymbol{N_{\widehat{+}}}$, and the "called" number of negatives, left column or $\boldsymbol{N_{\widehat{-}}}$. For example, Figure 1c shows an accuracy of 95% with TP of 27,561, TN of 26,267, FP 2104, and FN of 766. The AdaBoost classifier results in a reduced number of False Alarms (FP) and almost no Missed Detections (FN) as expected from Schapire and Singer (1999), Murphy (2012), and Al-Salemi et al. (2019) with a Sensitivity Rate of 97.3%, a Specificity Rate of 92.6%, a False Alarm Rate of only 7.4%, and a Missed Detection Rate of 2.7%. The Precision Rate of 92.9%, not shown in the confusion matrix because the denominator is the sum of the right column or $\boldsymbol{N_{\widehat{+}}}$, the predicted number of positives.

4.3. ROC curves

A further consideration is the extent of true positives and false negatives. The receiver operating characteristic (ROC) curve (Fawcett, 2006) is a means to assess this. Figure 4 uses the ROC plot: the Sensitivity Rate or True Positive Rate against the False Alarm or False Positive Rate at different thresholds $\boldsymbol{\tau}$ or cut-offs of the probability of being an Ent. If $\boldsymbol{\tau = 0}$ we are at the top right corner of Figure 4 where everyone is classified as an Ent so the True Positive and False Positive Rates both equal one as the $\boldsymbol{TP, FP > 0}$ while $\boldsymbol{FN, TN = 0}$. An analogous case, when $\boldsymbol{\tau = 1}$ is at the bottom left corner where everybody is classified as a W and both rates are now zero as $\boldsymbol{TP, FP = 0}$ while $\boldsymbol{FN, TN > 0}$. Along the diagonal and for different $\boldsymbol{\tau s}$ the two rates are equal as long as the Ent/W assignment is random. We plot the ROC curve for the following classifiers: RandomTrees (RT), RandomForest (RF), GradientBoosting (GBT) as both stand-alone methods and combined with LR following a code by Head (2018) under a 3-clause BSD License. The best classifier is the one which achieves the top left corner. Again, the preferred comparison classifier is a *boosting* method—as discussed in Friedman (2001) and James et al. (2013). In fact Gradient Boosting associated with the LR has the purple or outer-most curve giving the best ROC curve among all the classifiers, see Figure 4b. Similar to *Section 4.1*, the second place is achieved by the *ensemble* method RT that relies on the *wisdom of the crowd*, see Géron (2017).

4.4. Bag of words

A bag of words, according to Chollet (2018), is the result of a two-stage process. First, a tokenization or the breaking of text into units called tokens, and second, a vectorization or the association of numeric vectors with the generated tokens. The term bag "refers to the fact that [one is] dealing with a set of tokens rather than a list or sequence: the tokens have no specific order."[1] The data are the same maximum possible set of "extracted" Ents, but now the feature OccStrings is added to the same maximum "extracted" set. This is similar to many web-search algorithms (Kucukyilmaz et al., 2017) and title extraction (Hu et al., 2006), without using the semantic links between textual items used by Kastrati et al. (2019). This produces the confusion matrix in Figure 5 with an accuracy of more than 99% still using the AdaBoost parametrization

[1]As a hash in Perl. See Figure 6-2. A hash as a barrel of data. (Schwartz et al., 2008). Or an R list, a Python dictionary or even a C structure, see (Matloff, 2011)

suggested by Dawe (2018). This result shows the power of using the full occupational descriptor text in the form of a bag of words to solve this ML task.

4.5. Deep Learning

Deep Neural Networks (DNN) are an important advance in the art of ML (McCulloch and Pitts, 1943; Rosenblatt, 1958; Rumelhart et al., 1985) which is particularly valuable for complex textual classification (Abdi et al., 2019; Kastrati et al., 2019). As suggested by Chollet (2018), a good metaphor of DL is the uncrumpling of a complicated manyfold of data. For example, imagine two sheets of colored paper: one green for Ents and one purple for Ws. Put them one on top of the other and crumple them into a ball of paper. Now you cannot tell them apart. DL consists of chains of geometric operations (underlying *tensor operations*) to uncrumple this ball of paper in order to separate—that is to classify—the Ents green sheet from the Ws purple one; or "finding neat representations for complex, highly folded manifolds" (Chollet, 2018). This is interpreted by Kastrati et al. (2019) as integrating learning into an ontology based on the semantics in the text, or by Abdi et al. (2019) as allowing sentiment analysis, although our text descriptors are too brief and simplistic to utilize such complex approaches. Using deep learning, it is possible to produce the best performance for the problem at hand with an accuracy of 96% after transforming the features to tensors, coding categorical variables as 2D-tensors with normalization, and building a sequential neural network with two Dense layers of sixteen hidden units, a "relu" (rectified linear unit, or non-linearity) activation, plus a final layer with just one hidden unit and a "sigmoid" activation. Notice that this model does not use the power of the OccStrings analyzed in the previous section. So the improvements can only be attributable to the DL method. Figure 6 shows the loss and the accuracy of both the training and the validation sets. Overall the model performs well both in the training and, importantly, also in the validation sets. This implies that the model *generalizes* well, since it performs well on data it has never seen before. Also, it suggests that *overfitting* to the training data is not a problem for this model with its current *capacity* (or number of learnable parameters) and amount of training data. Figure 7 shows the output of the final layer as seen in TensorBoard (a suite to visualize learning and "make it easier to understand, debug, and optimize TensorFlow programs" (Abadi et al., 2015)). This is a bimodal distribution of probabilities with Ents as ones and Ws as zeros.

5. Assessment and Conclusion

This paper uses methodological advances in machine learning to apply to historical census data classification. In particular, it has shown that boosting, followed by ensemble methods, sometimes associated with LR, among the probabilistic approaches generate sizable improvements in accuracy over the benchmark of a stand-alone LR for the classification of individuals by their entrepreneurial status for the early censuses. The results tend to confirm Hindman (2015) who suggests that "Ensemble models illustrate what is possible in terms of predictive accuracy, and they provide the best yardstick with which to judge simpler models" ... "Ensemble methods .... are almost always a superior choice to the OLS and logit models that dominate empirical social science work today". However, significant improvements in accuracy of the census classifications assessed here can be achieved with a bag-of-words strategy using the OccString feature in the data, which employs advances in text and natural language preprocessing. At the same time, DL with neural networks, can also give substantial improvements over the traditional LR method. This confirms that ML and DL can be actively developed to tackle classification of historical data. This case study has proved that ML and, in particular, DL are techniques that are valuable for classification of historical data; they also encourage subsequent exploration of record linkage of historical census data as recently described by Capobianco and Marinai (2019) and Liu et al. (2019). Our efforts demonstrate that a multidisciplinary approach to traditional information classification tasks can realize the potential of a "big data revolution" (Hindman, 2015). As Hindman (2015) suggests "[n]ew data sources and better algorithms do allow social scientists of all stripes to offer most accurate forecasts in many ... areas". Finally, the addition of machine learning to traditional methodological techniques suggests that the use of big data techniques (even for small sample queries) can help to understand and improve testing of theory.

The paper has focused on the methodological advances offered by ML, and the comparison of different methods for implementing ML. Future developments of these methods can be used to join up the historical censuses with modern data so that the long term trends in entrepreneurship can be examined over time. This begins to allow evaluations of the effects of changing descriptors of entrepreneurial behaviour, and the effect of different economic conditions on decision choices between waged work and employer or own account status. Indeed the results of the application of the methods used here for identifying entrepreneurial status 1851-81 linked to the later period

1891-1911, and then linked to modern censuses show that the Victorian period had a higher rate of entrepreneurship than any subsequent time in Britain (Bennett et al., 2019d). In addition, the availability of a database on the full population of entrepreneurs over time can be used to study the statistical characteristics of the firm size distribution (Montebruno et al., 2019b,c), and the study of the determinants of Victorian entrepreneurship (Bennett et al., 2019a,d). Moreover the data deposit of the estimates of entrepreneurial status based on the methods used in this paper allow other researchers to develop answers to other research questions; such as persistence in entrepreneurship over time, growth and change in firm sizes, and using record-linkage between census years, open up new potential to examine the life stages and career evolution of entrepreneurs and switching between different employment statuses.

Acknowledgments

This research has been supported by the ESRC under project grant ES/M010953: **'Drivers of Entrepreneurship and Small Businesses'** and Isaac Newton Trust research grant 17.07(d): **'Business Employers in 1871'**.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. URL: `http://tensorflow.org/`.

Abdi, A., Shamsuddin, S.M., Hasan, S., Piran, J., 2019. Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. Information Processing and Management 56, 1245–1259.

Al-Salemi, B., Ayob, M., Kendall, G., Noah, S.A.M., 2019. Multi-label Arabic text categorization: A benchmark and baseline comparison of multi-label learning algorithms. Information Processing and Management 56, 212–227.

Alvarez-Galvez, J., 2016. Discovering complex interrelationships between socioeconomic status and health in Europe: A case study applying Bayesian Networks. Social Science Research 56, 133–143. URL: `http://www.sciencedirect.com/science/article/pii/S0049089X16000107`, doi:https://doi.org/10.1016/j.ssresearch.2015.12.011.

Bennett, R.J., Montebruno, P., Smith, H., van Lieshout, C., 2018. Reconstructing entrepreneur and business numbers for censuses 1851-81. Working paper 9. URL: `https://www.repository.cam.ac.uk/handle/1810/290508`, doi:https://doi.org/10.17863/CAM.37738.

Bennett, R.J., Montebruno, P., Smith, H., van Lieshout, C., 2019a. Entrepreneurial discrete choice: Modelling decisions between selfemployment, employer and worker status. Working paper 15. URL: `https://www.repository.cam.ac.uk/handle/1810/290086`, doi:https://doi.org/10.17863/CAM.37312.

Bennett, R.J., Montebruno, P., Smith, H., van Lieshout, C., 2019b. Reconstructing proprietor numbers for censuses 1851-81: Extension and alternative Working paper 9.2.

Bennett, R.J., Smith, H., van Lieshout, C., Montebruno, P., Newton, G., 2019c. British Business Census of Entrepreneurs 1851-1911 (BBCE), UK Data Archive data deposit. Forthcoming .

Bennett, R.J., Smith, H., van Lieshout, C., Montebruno, P., Newton, G., 2019d. The Age of Entrepreneurship: Business Proprietors, Self-employment and Corporations Since 1851. Routledge international studies in business history.

Blanchflower, D.G., Oswald, A.J., 1998. What Makes an Entrepreneur? Journal of Labor Economics 16, 26–60.

Boutell, M.R., Luo, J., Shen, X., Brown, C.M., 2004. Learning multi-label scene classification. Pattern Recognition 37, 1757–1771.

Cameron, A.C., Trivedi, P.K., 2005. Microeconometrics: methods and applications. Cambridge University Press, Cambridge.

Capobianco, S., Marinai, S., 2019. Deep neural networks for record counting in historical handwritten documents. Pattern Recognition Letters 119, 103–111. URL: `http://www.sciencedirect.com/science/article/pii/S0167865517303914`, doi:https://doi.org/10.1016/j.patrec.2017.10.023.

Cheng, W., Hüllermeier, E., 2009. Combining instance-based learning and logistic regression for multilabel classification. Machine Learning 76, 211–225.

Chollet, F., 2018. Deep learning with Python. Manning, Shelter Island, NY.

Dawe, N., 2018. Python Code for Two-class AdaBoost (3-clause BSD License).

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874. URL: `http://www.sciencedirect.com/science/article/pii/S016786550500303X`, doi:https://doi.org/10.1016/j.patrec.2005.10.010.

Freund, Y., Schapire, R.R.E., 1996. Experiments with a New Boosting Algorithm. International Conference on Machine Learning , 148–

156URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.6252, doi:10.1.1.133.1040, arXiv:978-0-387-09823-4_45.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29, 1189–1232.

Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., Brinker, K., 2008. Multilabel classification via calibrated label ranking. Machine Learning 73, 133–153.

Géron, A., 2017. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. First edit ed., O'Reilly, Sebastopol, CA.

Goldberger, A.S., 1991. A course in econometrics. Harvard University Press, Cambridge, Mass.; London.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. Adaptive computation and machine learning, The MIT Press, Cambridge, Mass.; London, England.

Head, T., 2018. Python Code for ROC curve (3-clause BSD License).

Higgs, E., 2004. Life, death and statistics: civil registration, censuses and the work of the General Register Office, 1836-1952. Local population studies supplement, Local Population Studies, Hatfield.

Higgs, E., Schürer, K., 2014. Integrated Census Microdata (I-CeM), 1851-1911, UK Data Archive data deposit SN-7481. URL: http://doi.org/10.5255/UKDA-SN-7481-1, doi:10.5255/UKDA-SN-7481-1.

Hindman, M., 2015. Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. The ANNALS of the American Academy of Political and Social Science 659, 48–62.

Hu, Y., Li, H., Cao, Y., Li, T., 2006. Automatic extraction of titles from general documents using machine learning. Information Processing & Management 42, 1276–1293. URL: http://search.proquest.com/docview/194922203/.

Hüllermeier, E., Fürnkranz, J., Cheng, W., Brinker, K., 2008. Label ranking by learning pairwise preferences. Artificial Intelligence 172, 1897–1916.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning: with applications in R. Springer texts in statistics.

Kastrati, Z., Imran, A.S., Yayilgan, S.Y., 2019. The impact of deep learning on document classification using semantically rich representations. Information Processing and Management 56, 1618–1632.

Kucukyilmaz, T., Cambazoglu, B.B., Aykanat, C., Baeza-Yates, R., 2017. A machine learning approach for result caching in web search engines. Information Processing and Management 53, 834–850.

van Lieshout, C., Bennett, R.J., Smith, H., 2019. Extracted data on employers and farmers compared with published tables in the Census General Reports, 1851-1881. Working paper 13.

Liu, Y., Jin, L., Lai, S., 2019. Automatic labeling of large amounts of handwritten characters with gate-guided dynamic deep learning. Pattern Recognition Letters 119, 94–102. URL: http://www.sciencedirect.com/science/article/pii/S0167865517303628, doi:https://doi.org/10.1016/j.patrec.2017.09.042.

Matloff, N.S., 2011. The art of R programming. No Starch Press., San Francisco, Calif.

McCulloch, W., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics 5, 115–133.

Montebruno, P., Bennett, R., Van Lieshout, C., Smith, H., Satchell, A., 2019a. Shifts in agrarian entrepreneurship in mid-Victorian England and Wales. The Agricultural History Review 67, 71–108.

Montebruno, P., Bennett, R.J., van Lieshout, C., Smith, H., 2019b. A tale of two tails: Do Power Law and Lognormal models fit firm-size distributions in the mid-Victorian era? Physica A: Statistical Mechanics and its Applications 523, 858–875. URL: http://www.sciencedirect.com/science/article/pii/S0378437119302079, doi:https://doi.org/10.1016/j.physa.2019.02.054.

Montebruno, P., Bennett, R.J., van Lieshout, C., Smith, H., 2019c. Research data supporting "A tale of two tails: Do Power Law and Lognormal models fit firm-size distributions in the mid-Victorian era?". Mendeley Data doi:http://dx.doi.org/10.17632/86xkkncmw3.1.

Murphy, K.P., 2012. Machine learning a probabilistic perspective. Adaptive computation and machine learning series, MIT Press, Cambridge, MA.

Murthy, D., Gross, A.J., 2017. Social media processes in disasters: Implications of emergent technology use. Social Science Research 63, 356–370. URL: http://www.sciencedirect.com/science/article/pii/S0049089X16306056, doi:https://doi.org/10.1016/j.ssresearch.2016.09.015.

Parker, S.C., 2004. The Economics of Self-Employment and Entrepreneurship. Cambridge University Press, Cambridge.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Rabe-Hesketh, S., Skrondal, A., 2012. Multilevel and longitudinal modelling using Stata. Volume 2, Categorical responses, counts, and survival. 3rd ed. ed., Stata Press, College Station, Tex.

Read, J., Pfahringer, B., Holmes, G., Frank, E., 2011. Classifier chains for multi-label classification. Machine Learning 85, 333–359.

Reichenberg, O., Berglund, T., 2019. “Stepping up or stepping down?“: The earnings differences associated with Swedish temporary workers’ employment sequences. Social Science Research URL: http://www.sciencedirect.com/science/article/pii/S0049089X18303156, doi:https://doi.org/10.1016/j.ssresearch.2019.04.007.

Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65, 386–408.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1985. Learning Internal Representations by Error Propagation. Technical Report. Institute For Cognitive Science, Univeristy of California, San Diego. La Jolla, California. URL: http://www.dtic.mil/dtic/tr/fulltext/u2/a164453.pdf.

Schapire, R., Singer, Y., 1999. Improved Boosting Algorithms Using Confidence-rated Predictions. Machine Learning 37, 297–336.

Schürer, K., Penkova, T., Shi, Y., 2015. Standardising and Coding Birthplace Strings and Occupational Titles in the British Censuses of 1851 to 1911. Historical Methods: A Journal of Quantitative and Interdisciplinary History 48, 195–213.

Schwartz, R.L., Phoenix, T., Foy, b.d., 2008. Learning Perl. 5th ed., O'Reilly, Beijing; Farnham.

Su, Z., Meng, T., 2016. Selective responsiveness: Online public demands and government responsiveness in authoritarian China. Social Science Research 59, 52–67. URL: http://www.sciencedirect.com/science/article/pii/S0049089X16302071, doi:https://doi.org/10.1016/j.ssresearch.2016.04.017.

Tang, B., Kay, S., He, H., 2016. Toward Optimal Feature Selection in Naive Bayes for Text Categorization. IEEE Transactions on Knowledge and Data Engineering 28, 2508–2521.

Tang, X., Chen, L., Cui, J., Wei, B., 2019. Knowledge representation learning with entity descriptions, hierarchical types, and textual relations. Information Processing and Management 56, 809–822.

The Editors of the American Heritage Dictionaries, 2011. The American Heritage dictionary of the English language. Houghton Mifflin Harcourt, Boston.

Tong, S., Chang, E., 2001. Support vector machine active learning for image retrieval, in: Proceedings of the ninth ACM international conference on multimedia, ACM. pp. 107–118.

Treasury Committee, 1890. Report of the Committee appointed by the Treasury to inquire into certain questions connected with the taking of the Census, presented to both Houses of Parliament by Command of Her Majesty. Minutes of evidence, appendices BPP 1890 LVIII. Technical Report. Printed for Her Majesty Stationery Office by Eyre and Spottiswoode. London.

Tsoumakas, G., Katakis, I., Vlahavas, I., 2011. Random k-Labelsets for Multilabel Classification. IEEE Transactions on Knowledge and Data Engineering 23, 1079–1089.

Varoquaux, G., Müller, A., 2018. Python Code for Classifier Comparison (3-clause BSD License). Modified for documentation by Jaques Grobler.

Wolpert, D.H., 1996. The Lack of A Priori Distinctions Between Learning Algorithms. Neural Computation 8, 1341–1390.

Wu, Q., Ye, Y., Zhang, H., Ng, M.K., Ho, S.S., 2014. ForesTexter: An efficient random forest algorithm for imbalanced text categorization. Knowledge-Based Systems 67, 105–116.

Zhang, M.L., Zhou, Z.H., 2007. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40, 2038–2048.

(a) 1891 Logistic Regression (1,000 W / 1,000 Ent)

(b) 1851 Logistic Regression (1,000 W / 1,000 Ent)

(c) 1851 AdaBoost (70,872 W / 70,872 Ent)

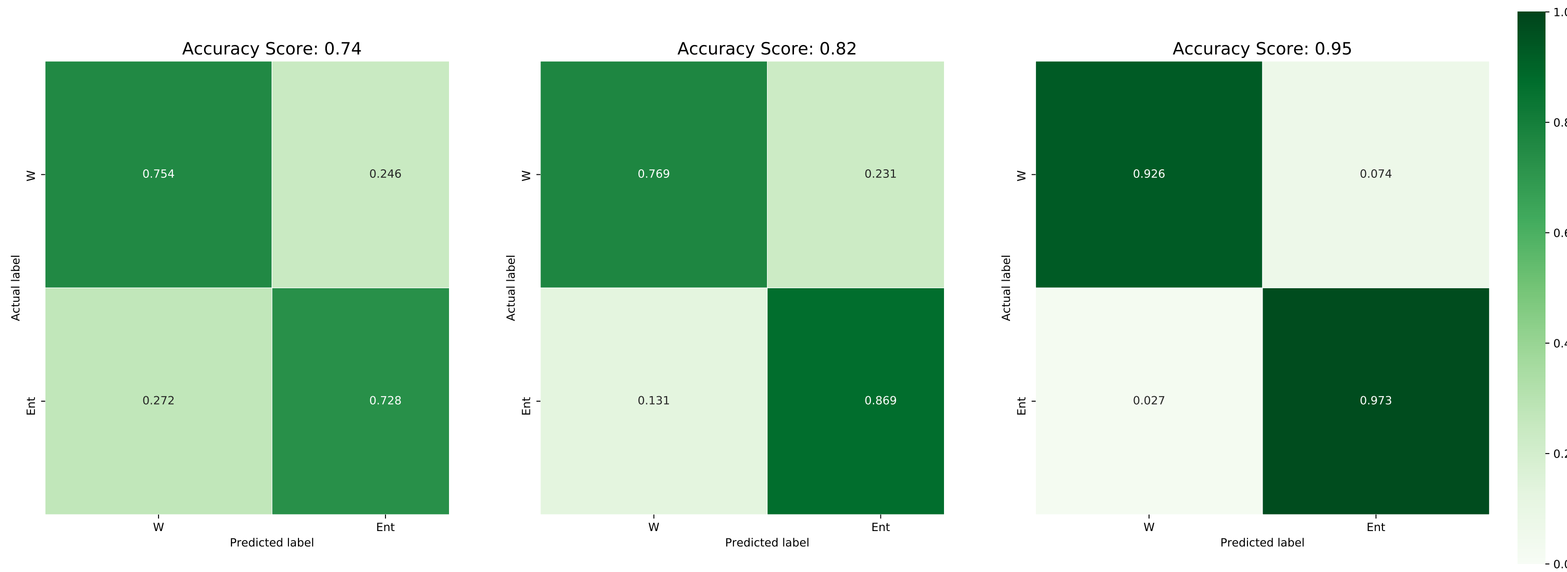


Figure 1: Confusion matrix for the binary classification of being a W or an Ent

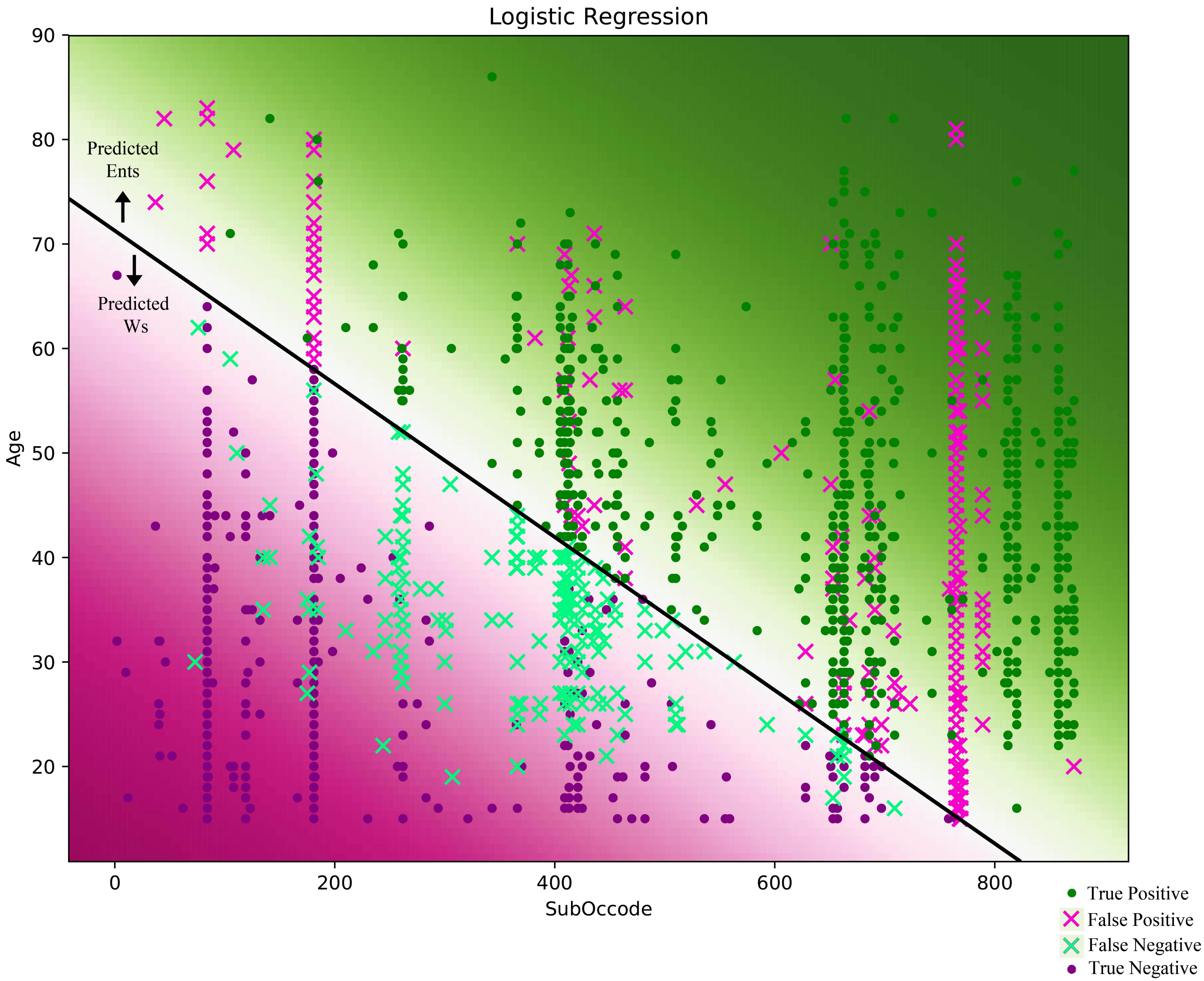


Figure 2: Ents and Ws classification: True Positive, False Positive, False Negative, True Negative

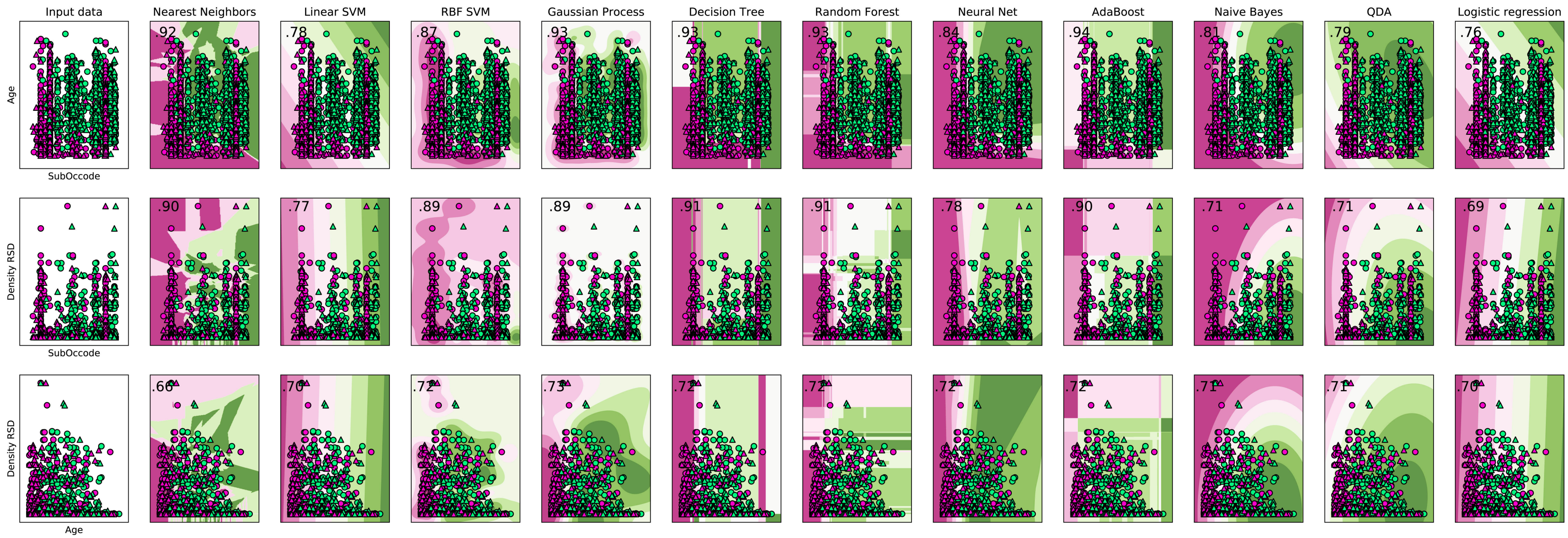
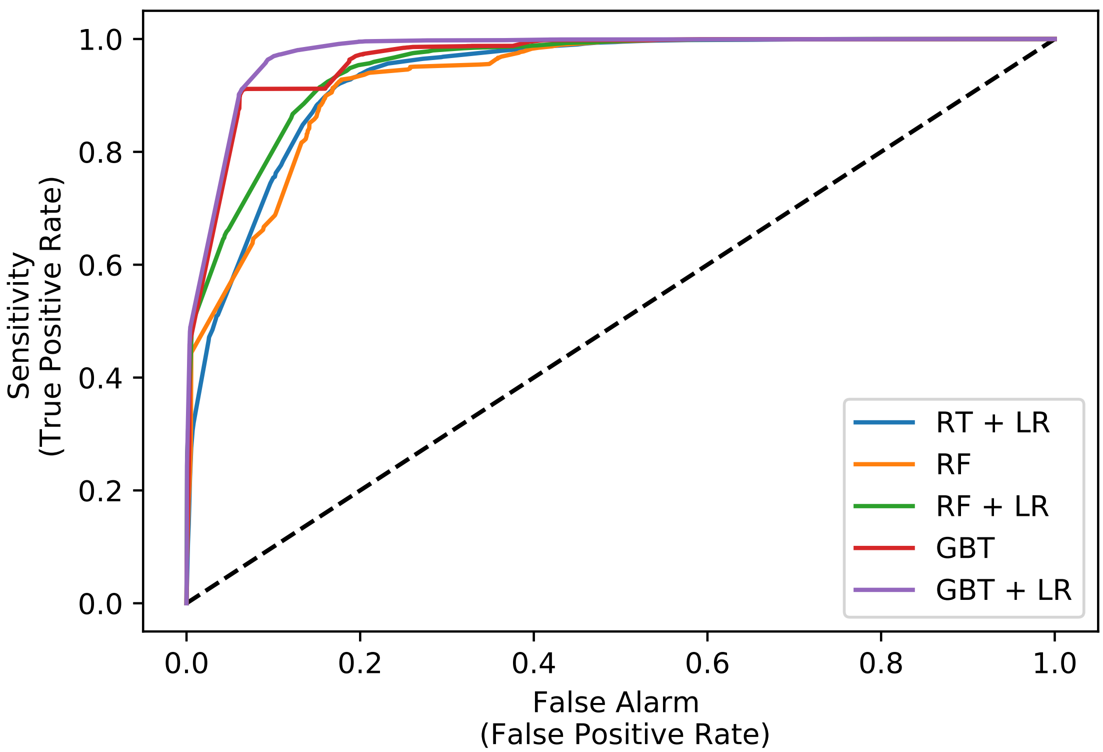


Figure 3: Accuracy and 2-D predicted probability grid comparing Logistic Regression (LR, far right) to ten alternative and competing classification algorithms (Nearest Neighbors, Linear SVM, RBF SVM, Gaussian Process, Decision Tree, Random Forest, Natural Net, AdaBoost, Naive Bayes, and QDA) for the label being an Ent with 2000 balanced random data points (1000 Ws, 500 Es and 500 OAs. That is 1000 Ws and 1000 Ents). The purple figures—that is circles and triangles—are Ws and the green ones are Ents. The circles are training set (60% of the total) and the triangles are the testing set (40% of the total). The color in the background is the 2-D predicted probability grid which means that when the grid is purple a test point—that is a triangle—is classified as W disrespecting of its true value or color and when the color is green a test point is classified as Ent, also disrespecting of its true value or color. According to this classification of the test points and accuracy for each method is being calculated. The code is used from (Varoquaux and Müller, 2018) under a 3-clause BSD License.

(a) ROC curve



(b) ROC curve (zoomed in at top left)

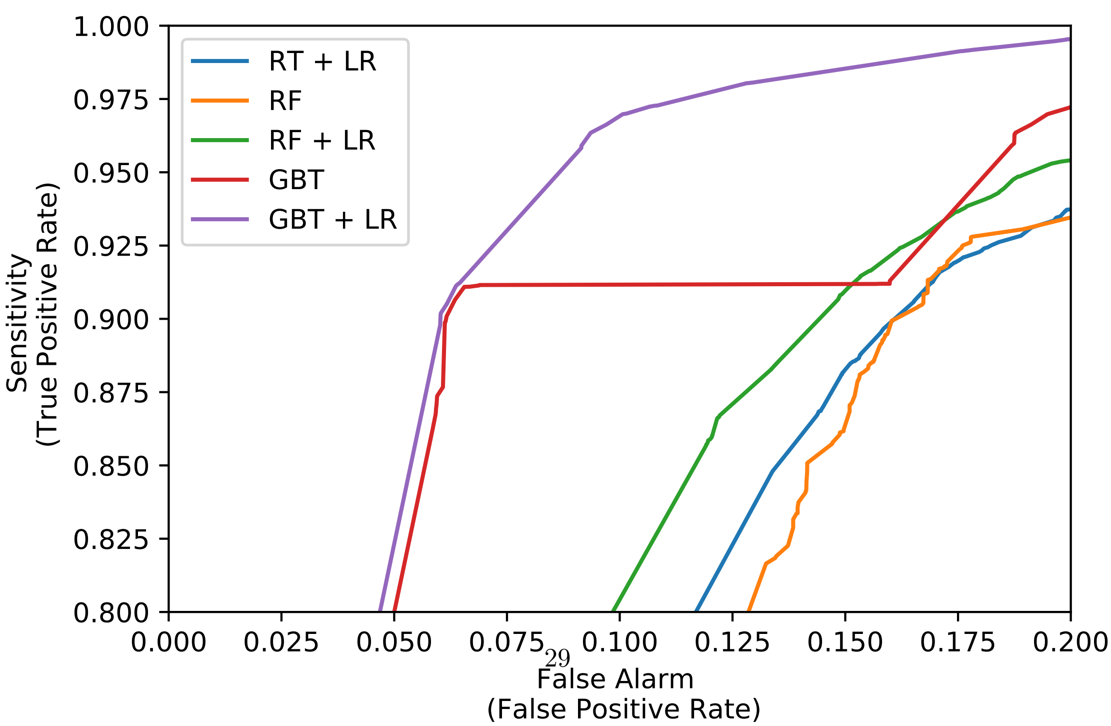


Figure 4: Receiving operating charateristic (ROC) curve for the binary classification of being or not an Entrepreneur using the following classifiers: RandomTrees (RT), RandomForest (RF), GradientBoosting (GBT) as stand-alone methods or combined with Logistic Regression (LR) (70,872 W / 70,872 Ent)

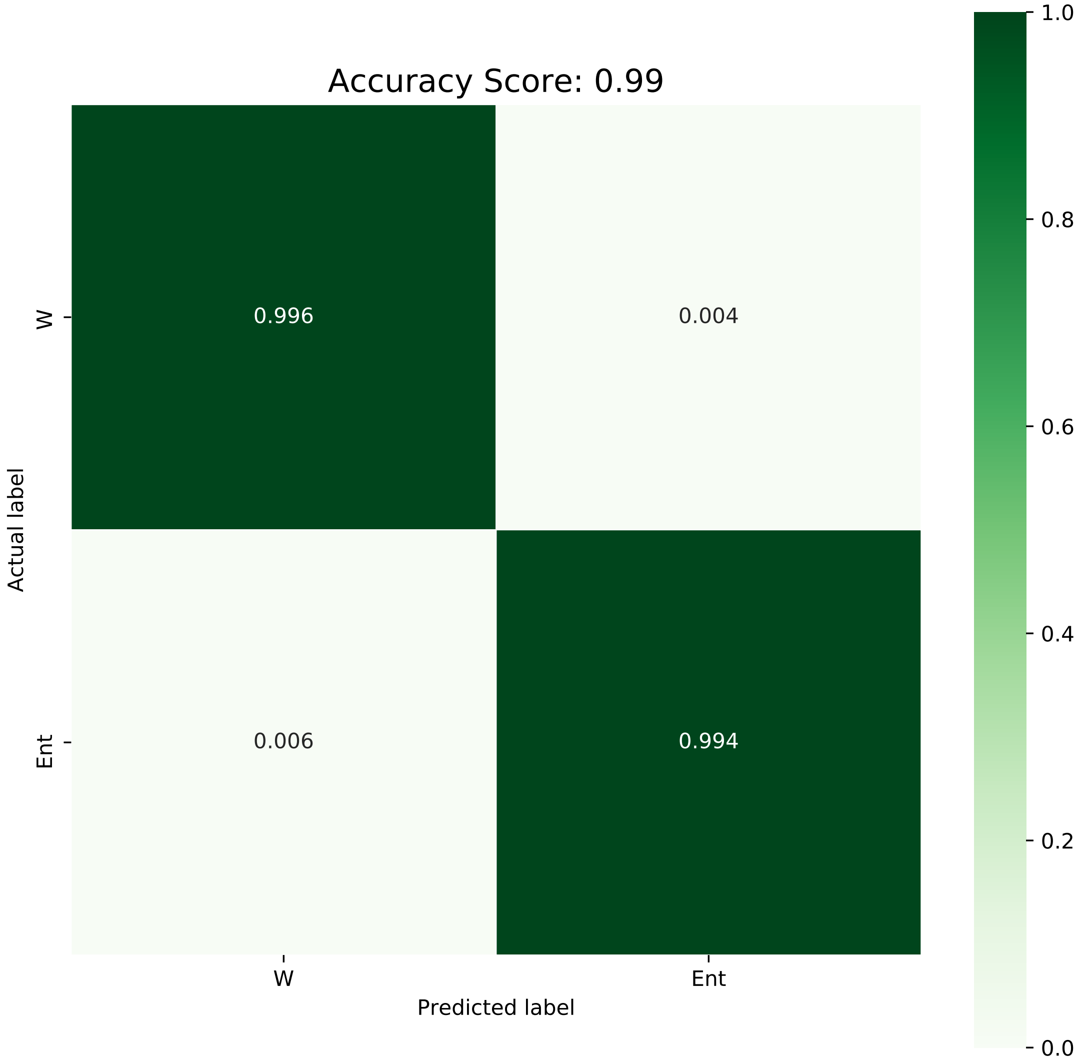
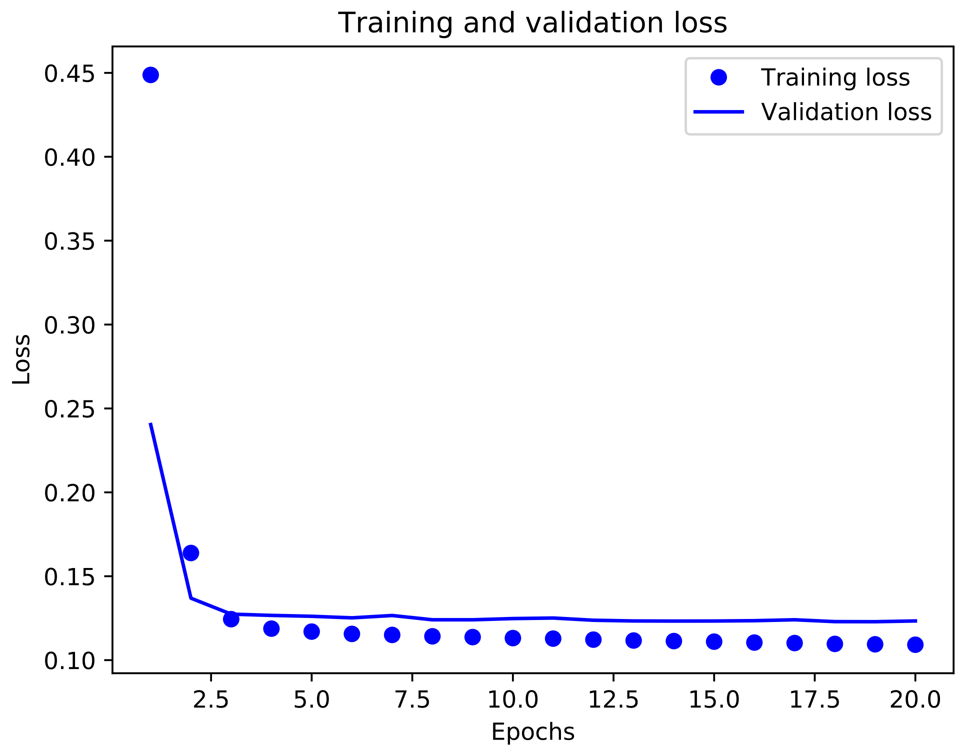


Figure 5: Confusion matrix for the binary classification of being a W or an Ent using the AdaBoost classifiers with the OccString feature (Bags of words), 1851 (70,872 W / 70,872 Ent).

(a) Loss



(b) Accuracy

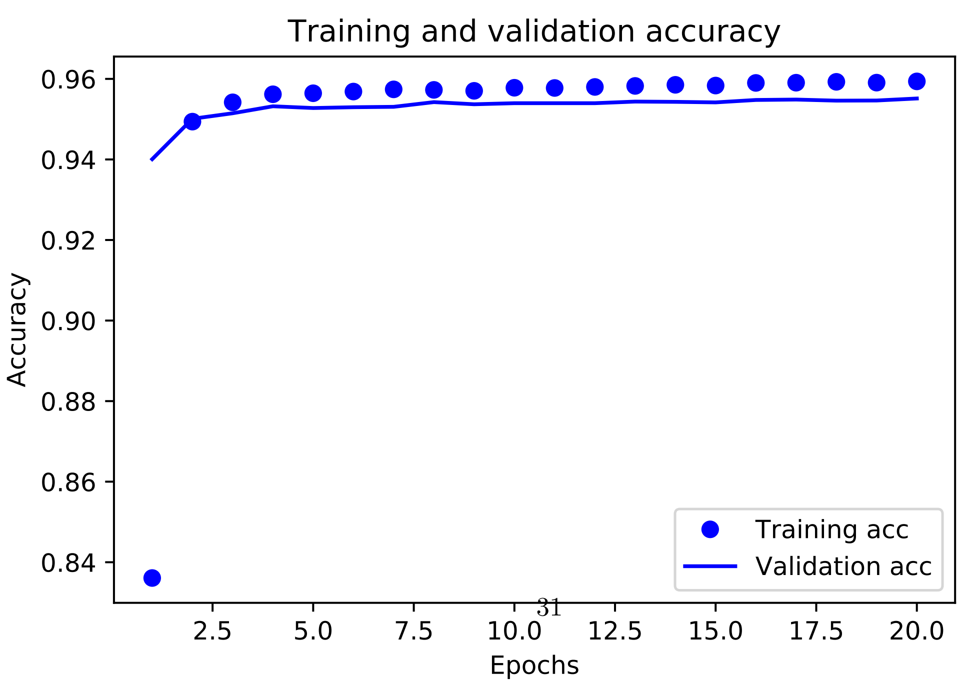


Figure 6: Loss and accuracy of a deep learning for the maximum possible set of extracted (that is labelled according to their strings or type of employment occupation codes) Entrepreneurs, after farmers have been dropped from the set; for 70,872 Workers, 35,436 Employers, and 35,436 Own accounts.

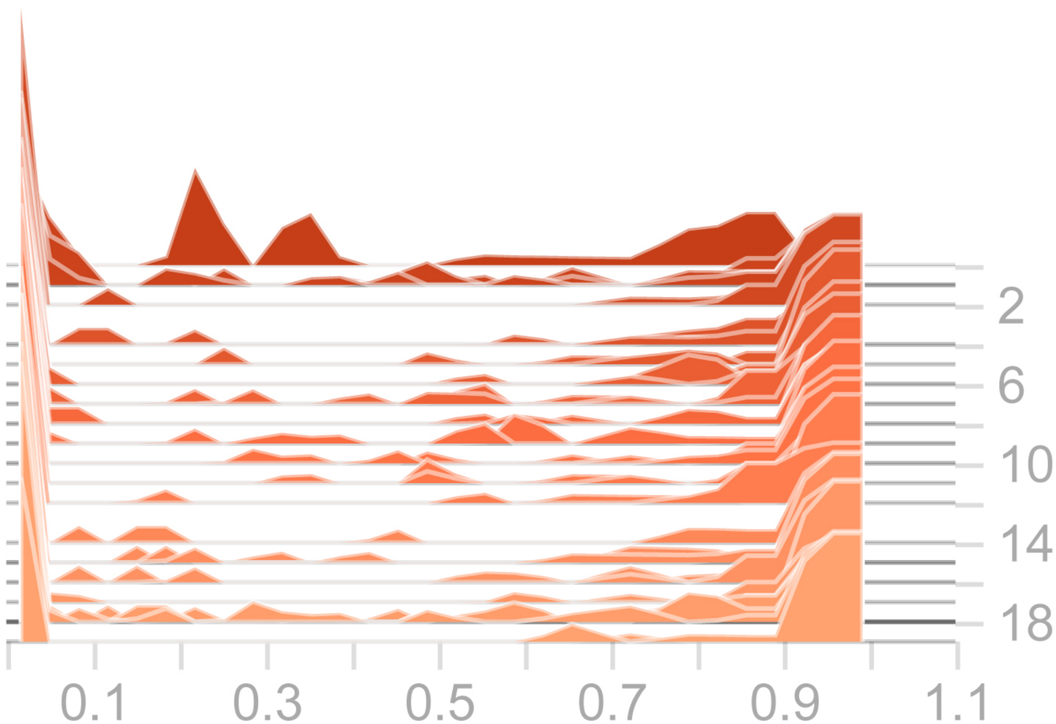


Figure 7: TensorBoard histogram of the output of the final layer with the "sigmoid" activation. It shows a bimodal distribution of Ents (1s) and Ws (0s). That is the scope of our classification effort: a probability of being an Ent.

	Means				Medians				Overall	
	Worker	Entrepreneur	t	p-value	Worker	Entrepreneur	z	p-value	min	max
Age	32.1	44.6	-976	0	28	42	-941	0	15	90
RSD Density	28.3	25.9	53.7	0	2.84	1.91	166	0	0	392
Sex	1.3	1.28	60	0	1	1	60	0	1	2
Marital status	1.57	2.07	-737	0	1	2	-767	0	1	4
Relationship to the head	2.97	1.61	627	0	2	1	813	0	1	10
Number of Servants	.0785	.432	-708	0	0	0	-847	0	0	99
SubOccode	344	487	-652	0	248	482	-589	0	1	882

Table 1: 1891 census training set features for binary labels: Worker (W) / Entrepreneur (Ent). The table shows the means, the t-value and p-value of the two-sample t-test with equal variances and the medians, the z-value and p-value of the two-sample Wilcoxon rank-sum, or Mann-Whitney, test for the the seven features of the training set. Sex is coded as 1 Male (base category), and 2 Female. Marital status as 1 Single (base category), 2 Married, 4 Widowed. Relationship to the head as 1 Head (base category), 2 Child and Family Unit (CFU) member, 3 Older generation, 4 Sibling, 5 Other family, 6 Servant, 7 Working title, 8 Lodger/boarder, 9 Non-household, 10 Unknown. RSD Density is Registration SubDistrict Density. SubOccode is coded 1 to 882 with SubOccode 196. Coal Miners - Hewers, Workers At The Coal Face as base category. (See Bennett et al. (2018) for a full list of the 882 categories of the feature SubOccode)

	Means								
				W/E		W/OA		E/OA	
	Worker	Employer	Own account	t	p-value	t	p-value	t	p-value
Age	32.1	45.9	43.9	-676	0	-753	0	79.1	0
RSD Density	28.3	22	28.1	89	0	3.5	.000445	-71.8	0
Sex	1.3	1.11	1.37	303	0	-142	0	-358	0
Marital status	1.57	2.07	2.08	-467	0	-607	0	-7.4	1.36e-13
Relationship to the head	2.97	1.34	1.75	459	0	450	0	-149	0
Number of Servants	.0785	.871	.184	-1,021	0	-207	0	428	0
SubOccode	344	435	517	-261	0	-646	0	-182	0

	Medians								
				W/E		W/OA		E/OA	
	Worker	Employer	Own account	z	p-value	z	p-value	t	p-value
Age	28	45	43	-653	0	-725	0	79.8	0
RSD Density	2.84	.866	2.26	211	0	58.4	0	-136	0
Sex	1	1	1	302	0	-142	0	-344	0
Marital status	1	2	2	-563	0	-573	0	65.2	0
Relationship to the head	2	1	1	635	0	568	0	-233	0
Number of Servants	0	0	0	-1,115	0	-334	0	443	0
SubOccode	248	409	657	-256	0	-566	0	-159	0

Table 2: 1891 census training set features for multi-class labels: Worker (W) / Employer (E) / Own account (OA). The table shows the means, the t-value and p-value of the two-sample t-test with equal variances and the medians, the z-value and p-value of the two-sample Wilcoxon rank-sum, or Mann-Whitney, test for the the seven features of the training set. Sex is coded as 1 Male (base category), and 2 Female. Marital status as 1 Single (base category), 2 Married, 4 Widowed. Relationship to the head as 1 Head (base category), 2 Child and Family Unit (CFU) member, 3 Older generation, 4 Sibling, 5 Other family, 6 Servant, 7 Working title, 8 Lodger/boarder, 9 Non-household, 10 Unknown. RSD Density is Registration SubDistrict Density. SubOccode is coded 1 to 882 with SubOccode 196. Coal Miners - Hewers, Workers At The Coal Face as base category. (See on-line only appendix for a full list of the 882 categories of the feature SubOccode)

Feature	Labels: W/Ent Weight (w)	t-stat
SubOccode		
52. Schoolmasters And Teachers (Default) Minus Suboccode 802	3.959***	(126.62)
105. Laundry Wrk: Washer, Iron, Etc. (Not Dom) Minus Suboccode 805	5.038***	(167.28)
141. Carmen Carriers Carters And Draymen	3.622***	(118.78)
173. Farmer, Grazier	7.342***	(242.39)
196. Coal Miners - Hewers, Workers At The Coal Face	0	(.)
262. Blacksmiths Minus Suboccode 812	3.861***	(127.46)
409. Carpenter, Joiner Minus Suboccode 820	3.601***	(120.29)
551. Cotton & Cotton Good Mf Weaving Processes	0.136*	(2.26)
653. Tailors Not Merchants- Default Minus Subocc 858	4.482***	(149.15)
657. Dressmakers	6.768***	(226.57)
663. Shoe & Boot Maker (& Repairer) Minus Suboccode 862	4.721***	(159.43)
RSD Density	-0.00706***	(-132.86)
RSD Density × RSD Density	0.0000175***	(79.98)
Age	0.135***	(239.21)
Age × Age	-0.00102***	(-165.09)
Sex		
1. Male	0	(.)
2. Female	-0.0363***	(-5.58)
Marital status		
1. Single	0	(.)
2. Married	-0.106***	(-17.45)
4. Widowed	-0.0167	(-1.95)
2. Female × 2. Married	0.297***	(33.05)
2. Female × 4. Widowed	0.0300**	(2.89)
Relationship to the head		
1. Head	0	(.)
2. CFU member	-0.829***	(-138.86)
3. Older generation	-0.893***	(-52.83)
4. Siblings	-0.728***	(-72.65)
5. Other family	-1.053***	(-77.18)
6. Servants	-3.186***	(-69.34)
7. Working title	-2.773***	(-75.91)
8. Lodgers/boarders	-1.184***	(-162.58)
9. Non-household	-1.429***	(-48.37)
10. Unknown	-0.574***	(-46.14)
Number of servants	0.524***	(154.61)
Constant	-8.851***	(-278.30)
Observations	7,213,217	

t statistics in parentheses
*** $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$**

Table 3: Logit model weights. Weights (w, second column) calculated from the 1891 training data for features (base category) SubOccode (196), RSD Density, Age, Sex (Male), Marital status (Single), Relationship to the head (Head), and Number of servants. Binary labels are Worker (W) and Entrepreneur (Ent). In parentheses, t-statistics (third column).

	Labels: W/Ent	
Feature	$\boldsymbol{w}$/se	$\boldsymbol{\partial y/\partial x}$/se
RSD Density	-0.005***	-0.001***
	(0.000)	(0.000)
RSD Density × RSD Density	0.000***	
	(0.000)	
Age	0.115***	0.009***
	(0.000)	(0.000)
Age × Age	-0.001***	
	(0.000)	
Sex		
1. Male	0.000	0.000
	(.)	(0.000)
2. Female	0.776***	0.143***
	(0.004)	(0.001)
Martial status		
1. Single	0.000	0.000
	(.)	(0.000)
2. Married	-0.286***	-0.036***
	(0.005)	(0.001)
4. Widowed	-0.268***	-0.039***
	(0.007)	(0.001)
2. Female × 2. Married	0.174***	
	(0.007)	
2. Female × 4. Widowed	0.011	
	(0.008)	
Relationship to the head		
1. Head	0.000	0.000
	(.)	(0.000)
2. CFU member	-0.902***	-0.138***
	(0.005)	(0.001)
3. Older generation	-1.128***	-0.162***
	(0.013)	(0.001)
4. Siblings	-0.807***	-0.127***
	(0.008)	(0.001)
5. Other family	-1.071***	-0.156***
	(0.013)	(0.001)
6. Servants	-3.240***	-0.253***
	(0.045)	(0.001)
7. Working title	-2.256***	-0.230***
	(0.035)	(0.001)
8. Lodgers/boarders	-1.208***	-0.169***
	(0.006)	(0.001)
9. Non-household	-1.549***	-0.195***
	(0.025)	(0.002)
10. Unknown	-0.232***	-0.043***
	(0.010)	(0.002)
Number of servants	1.018***	0.150***
	(0.003)	(0.000)
Constant	-4.132***	
	(0.010)	
Observations	7,213,217	
Pseudo R2	0.193	
Chi-squared	887,072.910	
p-value	0.000	
*** $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$**		

Table 4: Logit model weights and marginal effects without the SubOccode. Weights ($\boldsymbol{w}$, second column) and marginal effects ($\boldsymbol{\partial y/\partial x}$, third column) calculated from the 1891 training data for same features and labels of the previous tables but without SubOccode for computing efficiency. Notable see that the Female weight is now positive. In parentheses and below, standard errors (**se**).

Feature	Labels: W/E/OA Weight ($\boldsymbol{w}$)	*t*-stat	Weight ($\boldsymbol{w}$)	*t*-stat
	2.Employer		**3.Own account**	
RSD Density	-0.00745***	(-105.62)	-0.00365***	(-72.43)
RSD Density × RSD Density	0.0000217***	(80.07)	0.00000929***	(42.80)
Age	0.119***	(158.85)	0.114***	(229.08)
Age × Age	-0.000793***	(-101.43)	-0.000778***	(-145.02)
Sex				
1. Male	0	(.)	0	(.)
2. Female	-0.271***	(-26.55)	0.947***	(206.47)
Marital status				
1. Single	0	(.)	0	(.)
2. Married	0.0396***	(4.74)	-0.504***	(-89.45)
4. Widowed	-0.283***	(-26.28)	-0.290***	(-38.29)
2. Female × 2. Married	0.115***	(7.44)	0.370***	(48.25)
2. Female × 4. Widowed	0.439***	(30.54)	0.00476	(0.57)
Relationship to the head				
1. Head	0	(.)	0	(.)
2. CFU member	-1.029***	(-108.67)	-0.872***	(-161.54)
3. Older generation	-1.242***	(-50.43)	-1.087***	(-73.87)
4. Siblings	-0.845***	(-52.84)	-0.785***	(-89.18)
5. Other family	-1.464***	(-46.07)	-0.962***	(-75.13)
6. Servants	-3.173***	(-29.52)	-3.268***	(-67.07)
7. Working title	-2.162***	(-25.26)	-2.276***	(-60.20)
8. Lodgers/boarders	-1.480***	(-119.60)	-1.158***	(-171.50)
9. Non-household	-1.829***	(-35.02)	-1.493***	(-54.88)
10. Unknown	-0.171***	(-9.08)	-0.257***	(-22.05)
Number of servants	1.430***	(361.84)	0.705***	(212.89)
Constant	-5.535***	(-308.16)	-4.390***	(-384.13)
Observations		7,173,550		

t statistics in parentheses

*** $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$**

Table 5: MNL model weights without SubOccode. Weights ($\boldsymbol{w}$, second and fourth columns) calculated from the 1891 training data for the features without SubOccode for ease of computation.is now positive. Multi-class labels are Worker (W), Employer (E), and Own account (OA) (Worker is base category). In parentheses (third and fifth columns), ***t*-stat**).

Feature	**Labels: W/E/OA** $\partial y/\partial x$/se	$\partial y/\partial x$/se	$\partial y/\partial x$/se
RSD Density	0.001***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)
Age	-0.009***	0.002***	0.007***
	(0.000)	(0.000)	(0.000)
Sex			
1. Male	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
2. Female	-0.143***	-0.014***	0.157***
	(0.001)	(0.000)	(0.001)
Martial status			
1. Single	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
2. Married	0.042***	0.005***	-0.047***
	(0.001)	(0.000)	(0.001)
4. Widowed	0.038***	-0.004***	-0.034***
	(0.001)	(0.000)	(0.001)
Relationship to the head			
1. Head	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
2. CFU member	0.136***	-0.038***	-0.098***
	(0.001)	(0.000)	(0.001)
3. Older generation	0.157***	-0.043***	-0.114***
	(0.001)	(0.001)	(0.001)
4. Siblings	0.124***	-0.033***	-0.090***
	(0.001)	(0.001)	(0.001)
5. Other family	0.152***	-0.048***	-0.104***
	(0.001)	(0.001)	(0.001)
6. Servants	0.246***	-0.063***	-0.183***
	(0.001)	(0.000)	(0.001)
7. Working title	0.224***	-0.057***	-0.167***
	(0.001)	(0.001)	(0.001)
8. Lodgers/boarders	0.167***	-0.048***	-0.119***
	(0.001)	(0.000)	(0.001)
9. Non-household	0.192***	-0.053***	-0.139***
	(0.002)	(0.001)	(0.001)
10. Unknown	0.042***	-0.007***	-0.035***
	(0.002)	(0.001)	(0.002)
Number of servants	-0.126***	0.053***	0.073***
	(0.000)	(0.000)	(0.000)
Constant		-4.132***	
		(0.010)	
Observations		7,173,550	
Pseudo R2		0.185	
Chi-squared		995,776.903	
p-value		0.000	

*** $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$**

Table 6: MNL model marginal effects without the SubOccode. Marginal effects ($\partial \mathbf{y}/\partial \mathbf{x}$) calculated from the 1891 training data for same features and labels of the previous tables but without SubOccode for computing efficiency. Multi-class labels are Worker (W), Employer (E), and Own account (OA) In parentheses and below, standard errors (**se**).