Generic auxiliary tasks to learn features that help regularize the network using Auto-encoder

Vinoth Vincent Raj, 1802566
University of Essex

***Abstract*—Autoencoders play a basic role in supervised and unsupervised learning and in deep learning architectures for transfer learning and one of a kind tasks. For two main reasons, the reduction of dimensionality has been an important subject of analysis in academia. Firstly, large volume of data resulting in computing costs and becomes difficult to handle. Also, mapping from higher dimensional to lower dimensional data without losing much information removes redundant dimensions in our dataset. Also, impact of the data sizes on the trained model results in high variance in model performance . In this paper, an approach of how autoencoders will be used for three different dataset which were collected from different domains will be discussed along with data description, experiments, impact of data sizes, limitations, challenges and their results will be presented. The results of the stacked autoencoder will also be compared with a standard algorithm and its performances will be discussed.**

***Index Terms*—Autoencoders, Data Exploration, Neural Network, Stacking.**

I. Introduction

Dimensionality curse is a common engineering problem. Ineffective and unreasonable reduction of the dimensionality of the variable jeopardizes the efficiency of machine learning, the accuracy of pattern recognition and the efficiency of data mining while increasing the workload of measured data experiments to some extent. Its dimensionality must be reduced in order to handle such real - world data adequately. Reduction of dimensionality is the transformation of high - dimensional data into a significant representation of low dimensions. In particular, high - dimensional representations are generated when signals, processes, images or physical fields are sampled.

Mostly dimensionality reduction techniques has many advantages. Firstly, Visualization: For the projection of high-dimensional data onto 2D or 3D. Secondly, Compression of data: Effective storage and recovery. Lastly, Removal of noise: Positive effect on accuracy of query.The rationale for dimensional reduction can be defined as follows from a knowledge discovery perspective, the identification of a reduced set of characteristics that predict results can be very useful. For several learning algorithms, the number of features directly increases the training and/or classification time.Noisy or meaningless features may have the same impact on classification as predictive features, thus negatively affecting accuracy.

In this paper, three different dataset from different domains have been selected. The three different domains are banking, government and medicine.

In banking data, given numerous features it requires to predict the customer satisfaction. Dramatically increasing apparent customer satisfaction or decreasing dissatisfaction has been recognized as the number one management objective in the banking industry. It is vital to know about unhappy customers to create competitive leadership, especially in the after - sales sector. Therefore, before the service interaction ends, the identification of potentially unhappy customers is of great value to enable proactive actions before the customer is actually unhappy. A completely satisfied consumer is retained and therefore from a business point of view is of great interest.

In the census data, the requirement is to correctly classify peoples income above or below a fixed threshold value based on numerous features like education, age, marriage status and location In recent years, the issue of income inequality has been of great concern. Better treatment of the destitute does not seem to be the only requirements for eliminating this problem. People in America sincerely believe that economic inequality is inexcusable and requires a good share of society's wealth. Governments in different parts of the world use different treatments to tackle economic inequality, although some success. Perhaps one of the main reasons for catastrophe is to do many things that reduce productivity and result in reduced results.

In medicine, ML techniques have been popularly used in intelligent health systems in recent decades, especially in the treatment and prognosis of breast cancer . Generally, a patient's diagnostic reliability depends on the experience of a doctor The erudite healthcare system can help doctors diagnose patients more accurately or even provide more tangible baselines and help people plan their physical condition in the future. BC has always had a high incidence rate and mortality rate as the most common cancer in women. BC alone is expected to account for 25% of all new cancer diagnoses and 15% of all cancer deaths among women worldwide, according to the latest cancer statistics [5] [6]. A precise classification can also help clinicians to prescribe the most appropriate treatment scheme. Classification is a complicated problem of optimization. Researchers have used many ML techniques to solve this classification problem.

II. Background

The autoencoder algorithm [2] is part of a special family of methods for reducing dimensionality using artificial neural networks. An autoencoder can be represented as a neural network whose primary function is to learn the underlying collector or the data set's feature space. An autoencoder tries to recreate the output inputs. It strives to try and learn a reduced structure of an input by reducing its error in reconstruction. The autoencoder algorithm and its extensions [6] [10] [1]

recently demonstrated a successful ability to learn useful data features that could lead to an "intrinsic data structure".

An autoencoder learns to squash data into a short code from the input neural network layer and then uncompress the data into something that closely represents the raw data. This enables the autoencoder to reduce dimensionality by trying to learn to ignore noise, for example.The encoding layer encodes the entire image into a matching code. The decoding layers will start to learn to decode the depiction of the learned code as closely as possible in its original form. More recently, autoencoders have once again taken center stage in the "deep architecture" approach,in which autoencoders, especially in the form of restricted Boltzmann machines, are stacked and trained in an unsupervised manner, followed by a supervised algorithm to train the encoded layer. Forecasting is an effective method to predicting bank clients ' level of customer satisfaction. The models documented in the research literature for bank client satisfaction are labeled into two main groups. Statistical models such as structural equation models and regression models are included [7][8][9][10] in the first category. There are some examples in this category. The main downside for this category of designs is the lack of decent predictive accuracy, now since the models have a linear structure, while customer satisfaction has a highly non-linear trend with the factors that influence them.

Artificial intelligence models like artificial neural networks (ANN) are included in the second group. ANN is an efficient process for modeling the actions of nonlinear structures with a simple system that makes them computationally efficient. At 15:51 2nd August 2015 (PT) telecommunications and healthcare services [12][13] ANN was used to anticipate customer satisfaction in non-banking organizations such as the automotive industry. No study of ANN modeling of bank customer satisfaction has been conducted in the literature. This study develops the first case to the best of the authors ' knowledge.

A recent analysis [14] mentions a generalized autoencoder (GAE), which concentrates on altering autoencoders to take into account the co relation between data by introducing a weighted relational function. In particular, the weighted distances between remodelled and unique instances are minimized. Even though ensuing applications[15][16] verify that the actual data link can enhance autoencoder overall performance in the decrease of dimensionality, the Generalized Autoencoder model seems to have some disadvantages. The current rebirth of interest in autoencoders is due to the advancement of training of deep architectures, since traditional gradient-based optimization techniques are are are are are not efficient when hidden layers are stacked with non-linearities several times. In 2006, Hinton[17] trained a deep network architecture empirically by sequentially optimizing each layer's weights in a RBM (Restricted Boltzman machine. Bengio succeeded in training a packed autoencoder with a comparable greedy layer approach on the MNIST dataset. This training approach addresses the issue of non-convex optimization that prevents deep network structure. Later studies show that the stacked autoencoder algorithm can learn substantive elements and therefore perform better in high dimensional data for classification.

Autoencoder is becoming increasingly popular as the training efficiency improves. It was soon found that the weights of neural network layers layers rise steeply due to matrix multiplication as layers are stacked, and then the imports of these weights also become bigger than the previous input characteristics. This overfitting problem leads to the fact that depictions of deep layers are much more likely to rely on the network topology than on the original input functions. Regularization is added in the autoencoder loss function to impose a penalty on large weights was added by Goodfellow et al.[19] Vincent et al.[20] advocated a Denoising Autoencoder (DAE) to resolve this problem by introducing input noises. The allocation of neural network hidden layer representations is not regulated by past studies. Variational Autoencoder[21] is therefore proposed to produce required representation allocations in neural network hidden layers. Autoencoders can be refereed as an unsupervised process to reduce dimensions and the decreased high-level features typically produce the critical insights from the source data.Hence, This does not make autoencoders susceptible to minor changes.

Contractive Autoencoder to make them resilient to minor variations was proposed by Rifai et al. [2]. Many GAE applications [23] prove that the maintenance of data relationships can produce good results, but outcomes depend heavily upon how distance weights are defined and selected. Due to the assignment of high weights to chosen relationships, the technique of robust, pre-defined distance weights is very arbitrary and can be biased in converting GAE into a supervised model. We therefore advocate a Relation Autoencoder (RAE) to reduce dimensionality by minimizing both the total loss of data characteristics.

III. Methodology

For this research purpose, three datasets have been chosen as discussed already. This section will explain the about the data in brief and also explore the data with the intuitive data analysis. The major results of the analysis will also be discussed in this section.

A. Census data

The information for our research were accessed from the Machine Learning Repository of the University of California Irvine (UCI). Barry Becker actually extracted it from the 1994 census database. The data set contains data from 48,842 unique records and 14 features for 42 countries. There are 14 features consisting of 6 continuous and 8 categorical features containing information on age, education, nationality, family status, employment, employment, gender, race, working hours per week, capital loss and capital gain, as shown in Table 1. The binary label in the data set is the level of income of that predicts whether or not a person earns more than 50 thousand dollars per year on the basis of the given set of features. Numerical column values have no missing values in

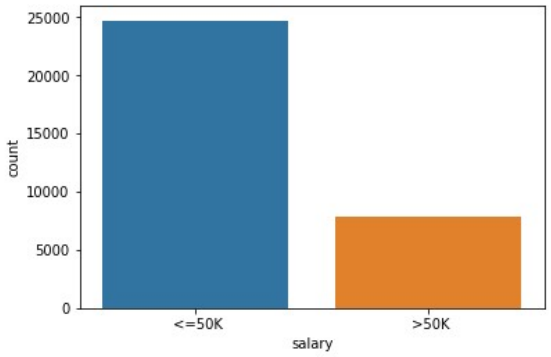

Fig. 1. Count of Salary feature which is the target to predict for the adult census data.

the dataset. The scales of each column are different. Hence it is advised to do a data normalization before working on the dataset.

B. Breast Cancer

Machine learning algorithms have been trained to recognize breast cancer by using Wisconsin Diagnostic Breast Cancer data (WDBC) [26]. The data consists of variables calculated from a digitalised picture of a fine needle aspirate (FNA) of a breast mass, according to[26].

The data source contains 569 data points with 31 features: 357-Benign and 212-Malignant, Benign instances are regarded as non - cancerous, i.e. non - life intimidating. However ,it could become a cancer status on a couple of occasions. An immune system termed ”sac” ordinarily separates benign tumors from many other cells and can be easily removed from the body.Malignant cancer begins with an unusual cell death and can actually spread or invade adjacent tissue quickly. The nuclei of malignant cartilage are usually much larger than in normal cells, which in possible future phases can be life - threatening.

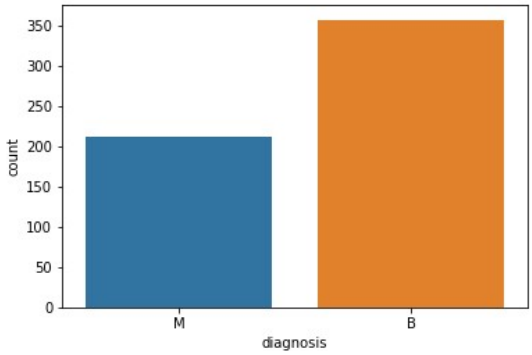

Fig. 2. Diagnosis values count for Malignant and Benign.

Some of the key features are radius, texture, compactness, concavity, perimeter, symmetry, smoothness, concave points, and fractal dimension. Each feature has three relevant information[26]: (1) average, (2) standard error, and (3) ”worst” or largest (mean of error).

C. Santander Customer Satisfaction

In this study, customer satisfaction prediction is the objective in of Santander Bank, a large company focuses primarily on the northeastern United States market Through a Kaggle competition (Santander, 2015), the aim is to find a suitable model to predict whether a certain customer will be unhappy with other qualities in the future. The introduction of this model can ensure that Santander is assertive.

The dataset consists of 371 anonymous columns and 76020 instances. The target variable of whether a customer is satisfied defined as a binary outcome in target column. The target is highly imbalanced as shown in the figure with 96% belonging to one class and the rest in other class. Based on some initial analysis we could find that unhappy customers have less products with the bank. Since the variables are anonymous it will be useful to plot feature importance from any model to get the variables contributing more to the target variable. Usually, in banking terms, it may be mortage value of the customer in the bank or the value of the customer The target

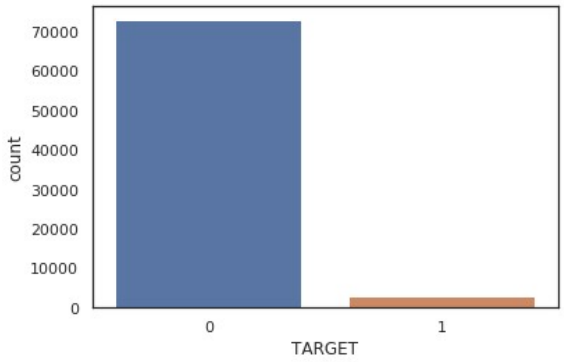

Fig. 3. Customer satisfaction count for santander data

variable is highly imbalanced as we could see in the figure above. Introducing SMOTE technique to do a oversampling could help the data to sample the target values equally which would be helpful for the machine learning algorithm to learn.

D. Data Pre-processing methods

Data pre-processing techniques includes data preparation, and data transformation into a suitable form for mining purposes. The goal of data pre-processing is to reduce the dimensions of the data, clean the data, find patterns in the data, normalize data and remove outliers. Since we are dealing with three sets of data. It is highly obvious that the features have different scales and each features can be ordinal, categorical and numerical. Also, to process any data it has to be converted to numerical format for the algorithms to access. The following pre-processing methods have been applied to comparatively to all datasets during this research

Data Cleaning: Some instances in the data had incomplete values, noise and outliers. The first step in data pre-processing is to correct the inconsistent data as these inconsistent data might affect the performance of our algorithm. Therefore, the data has been cleaned before passing it to algorithm

Converting categorical (text) values into numerical values : Many of the variables in Census data are categorical in which are in text. The categorical variables are converted into numeric values.

Removing unrelated columns: Most of the variables in Santander customer prediction were removed due to unrelated. The variable which is not contributing to the target have been removed using various feature selection techniques

Plot correlation matrix: The next technique which could be implemented is the correlation plot across all the variables to gain information about the correlation between each variable and the target variable also the correlation between independent variables.

Label encoding: Since our dataset contains text as values in some of the rows. Especially in breast cancer dataset the target value has two categories in text. Label encoder encodes the values between 0 to 1 for two classes and it depends on the number of classes a feature has. The drawback of label encoding is that it gives weight for higher numbers and may derive false assumptions.

Data Upsampling SMOTE: SMOTE, an over-sampling technique, is used in this experiment. SMOTE,is a popular over-sampling technique that balances class distribution by synthetically generating new minority class cases in the path of their nearest neighbors from current minority class instances. SMOTE aims to create new instances of a minority class close to borderlines to establish class boundaries.

Normalization: Normalization is the adjustment of values calculated on different scales, often before averaging, to a common scale. Normalization can refer to more robust modifications in more complex cases in which the motive is to align the entire probability distribution of normalized values. all the three datasets have been normalized before passing into the autoencoder.

IV. Experiments

For all the three datasets, an individual autoencoders will be trained first to find the best parameter for the autoencoder algorithm to replicate out input data. Then the encoder part which is the compressed learned features of all the input will stacked with a neural network to predict the classes of the target. But this steps will be experimented by taking smaller to all data points in the dataset. Their performace will be measured for different sample proportions and the autoencoder bottleneck neuron counts.The following sppraches will be taken. Firstly training a autoencoder and passing the well trained auto encoder as input to an artificial neural network and predict the target variable.

A. Training Auto encoder

An autoencoder is a neural network that learns efficient data coding by replicating the structure of its input. The objective of an autoencoder is to try to learn to comprise a set of data, typically to reduce dimensionality, by training the network to ignore the signal of ”noise.”

A greedy layer-wise training is a great way to get a stacked autoencoder decent parameters. Train the first layer of unprocessed input to obtain parameters W (1,1), W (1,2), b (1,1), b (1,2). To turn the raw input into a vector that activates the hidden units, use the first layer. Train the next layer of this vector for parameters W (2,1), W (2,2), b (2,1), b (2,2). Repeat for consequent layers using each layer's output as an input for the further layer. This technique trains each layer's parameters separately while making the remaining model parameters constant. To produce good outcomes, fine-tuning u after this training phase has been completed. This

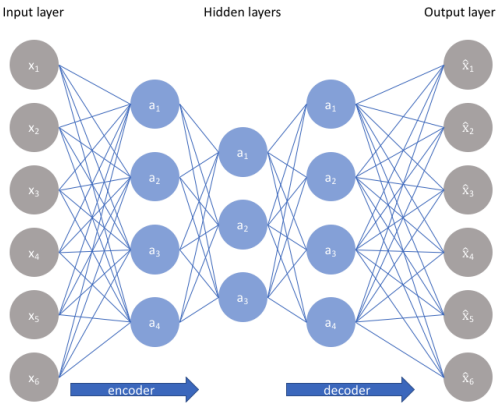

Fig. 4. Architecture of a deep autoencoder.

type of network consists of two parts:

Encoder: Encoder part of the network deforms the input into a depiction of latent space. The encoding functionality h= f (x) can represent it.

Decoder: Decoder segment is intended to recreate the input from the real representation of latent space. A decoding function r= g (h) can represent it.

Autoencoders are expected to keep as many information as possible whenever an input runs through the encoder and then the decoder, but they are still trained to have different nice properties in the new display. It is a feed-forward neural network that aims to minimize the loss function from the output which is the difference between the actual input and the output Once the loss is minimized it means that once autoencoder has been trained.This implies that once properly trained autoencoders are quite accurate and it will be hard to generalize data sets other than those for which they have been trained

B. Stacked Auto-Encoder

Once the autoencoder has been trained to replicate the exact data with minimum loss. The encoder pat can be passed as an input to the neural network. This functionality is basically a dimensionality reduction technique in which the encoder compress all the information into bottle neck neurons. These neurons have the compressed input information for all the data.

This neurons will now be passed into a traditional artificial neural network to predict their targets as shown in the figure below.

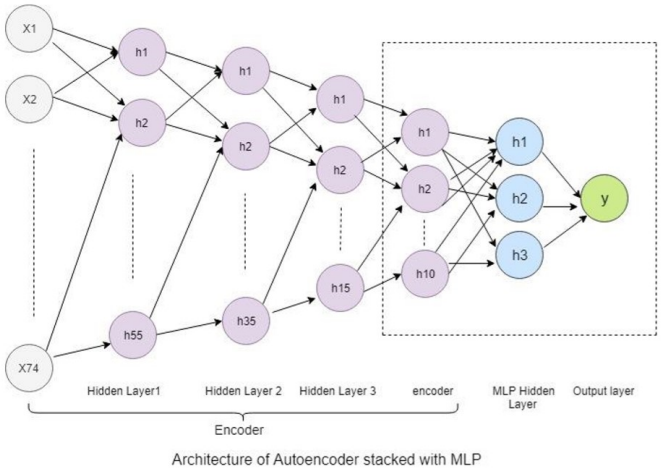

Fig. 5. Example architecture of a Stacked autoencoder with a deep neural network.

As shown in the the figure above, the hidden layers are trained by an unsupervised algorithm and then adjusted by a supervised method. Stacked autoencoder largely consists of three steps. Use raw data to train and obtain trained data. The data trained from the previous layer is used as an input for another layer until the training has been completed. To minimize loss and update weights with the training set, use the back propagation algorithm to achieve fine tuning once all hidden layers are trained. The recent advancements in Stacked Autoencoder gives a data version with much accurate and detailed and and encouraging variable information that is used to train a algorithm in a specific context and find decent accuracy than raw data training.

V. RESULTS

After pre-processing all the three datasets, dimensionality reduction has been done using autoencoder. The study was precise in compressing the dimensions to its half and check if the model is able to compete with other state of the art algorithms.As discussed in the last section the compressed features will now be trained using a artificial neural network and its performance will be tested. Understanding the impact of data size is imperative to make critical any business decisions. Hence throughout this process, data will be passed into the network on batches starting from 20% to 100%. At each of the batches, the model's performance will be tracked. Evaluation metrics like precision, recall, AUC,ROC score will be computed for each batches.Then a comparative study will be made discussing the results of our model. The structure, rationale, hyper parameter tuning will be discussed in the upcoming sections for each dataset.

a) Breast Cancer:
Breast cancer is the smallest dataset chosen for this study. Over this data, different autoencoder structures have been implemented by tweaking the algorithm parameters. A structure which was performing well for all the batch sizes were chosen. The breast cancer data has 30 features and as per the study objective the dimension has been reduced to 4. The Auto encoder architecture used was 30,8,4,8,30 where 4 is the bottle neck neuron count where the important features have been compressed.Autoencoders are instantly learned from instances of data. It implies it is trainable specialized neural network cases that will execute on a particular type of input and require no current engineering, only the appropriate training data. The other architectures which resulted in more loss are 30,2,30 and 30,10,30. The least loss was obtained in the last batch (100%) with 0.0049.

A trend could be observed in the loss for autoencoder where it finds it hard to converge processing the first batch but as the algorithms starts learning a quick and sharp convergence could be observed in the other iterations. This is due to the effect of learning rate in the algorithm. When training neural networks, batch size regulates the accuracy of estimating the error gradient. The main functions of the learning algorithm are batch, stochastic, and minibatch gradient descent. There is a relation between the batch size and the learning process's speed and stability. Different batch sizes have been implemented in the auto encoder. However, due to size of the data being small, a batch size of 4 have been used.Since Sigmoid and tanh functions are not appropriate for hidden layers as if z is very large or very small, the function slope becomes very small, slowing down the gradient descent process. RELU has bee used throughout the network with a sigmoid function at the output layer as the data has been normalized between 0 to 1. 10 epochs were chosen to avoid overfitting.

The compressed neurons are then taken and passed into a stacked autoencoder with 30 neurons in the hidden layer and one neuron at the output layer with sigmoid function generating the probabilities between 0 and 1. Later these probabilities are converted into binary classes by applying a threshold greater than 0.5. The results of the model could be found in the table below. A steady increase in accuracy has been observed as the data size increases. Precision and recall were also high when predicting with the full dataset.

Sample Size	Accuracy	AUC	F1 Score	Recall	Precision
20%	0.792	0.81	0.79	0.79	0.86
40%	0.905	0.91	0.91	0.91	0.92
60%	0.958	0.96	0.96	0.96	0.96
80%	0.947	0.95	0.95	0.95	0.95
100%	0.966	0.97	0.97	0.97	0.97

TABLE I
RESULTS OF STACKED AUTO ENCODER FOR EVERY BATCH SIZE FOR BREAST CANCER DATA

A steady increase in accuracy can be observed as sample size increases. In order to compare the performance of the stacked auto encoder model, another three different algorithms

were tested on the dataset in order to check the maximum accuracy it can obtain. KNN,SVM and Logistic Regression were implemented.SVMs do not provide specific calculations of probability, which are measured using a costly five-fold cross-validation Logistic regression outperformed other algorithms and achieved an 96& accuracy. The results could be seen in the below table.

Model	Precision	Recall	F1 Score
Logistic Regression	0.96	096	0.96
Support Vector machine	0.93	0.93	0.93
K-Nearest Neighbours	0.93	0.93	0.93

TABLE II
RESULTS OF OTHER ALGORITHMS ON BREAST CANCER DATASET.

It could be observed that the stacked autoencoder model performed equally well when comparing with other algorithms. The values of the results shown should be observed with caution as it may change based on the seeds. Hence, it is not possible to confirm that the stacked autoencoder model outperformed other state of the arts models. However, achieving the same target with limited features results in less model complexity and computationally inexpensive to train models over the entire data. This example demonstrated a possibility of achieving results that compete with other standard algorithms.

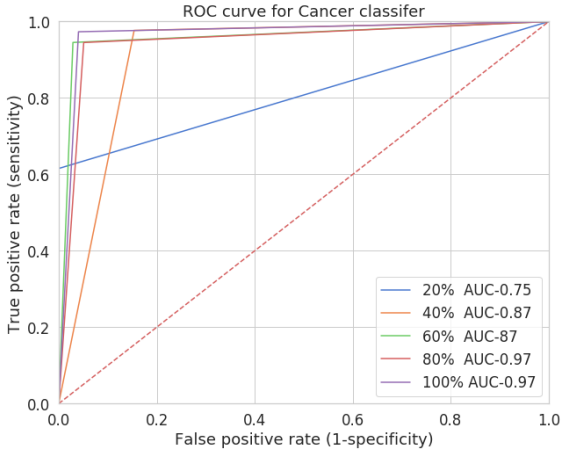

Fig. 6. ROC curve for Breast Cancer data.

The operating characteristic of the receiver (ROC) curve is a chart for a binary classifier to plot sensitivity vs. specificity, where different threshold values are compared. In our scenario, sensitivity is the likelihood that predicted statement is true. Specificity, however, is legitimate the likelihood of predicting classgiven the true state[27]. The ROC curve is usually presented by plotting false positives (FP) versus true positive (TP) In the figure abovewe compare the area for all classifiers under the ROC curve (AUC). The maximum value for AUC was obtained for 80% and 100% data.

b) Census data:

Census data is a medium sized data with 32561 instances but it is quite large when compared to Breast cancer dataset.The preprocessed data for Census is passed into different auto encoder structures that gives the minimizes the validation and training loss. Autoencoder structure of 15,10,5,10,15 has been chosen which obtains the least loss with a bottle neck of 5 neurons from 15 features. Different structures were tried by tweaking the parameters in the code. Batch size of 64 and 128 has been selected for autoencoder and stacked autoencoder with MLP. A batch size of 128 is used in updating the system weights, 128 samples from the train data will be used to calculate the error gradient. One epoch implies that one run through the training dataset was made by the learning algorithm, where instances were divided into randomly chosen ”batch size” groups. Optimal batch size were selected by trail and error basis and based on the generalization of the accuracy of the model. 10 epochs were given for both the autoencoder and stacked autoencoder with MLP as over fitting and fluctuating of the accuracy have been observed for higher epochs. Similar to breast cancer dataset, RELU activation function has been chosen for hidden layers and sigmoid activation for the output layer.
Adam optimizer has been used to train the stacked model as it is computationally efficient and requires less memory. Adam optimizer is widely known for learning rate decay as the direction towards global minimum becomes steeper learning rate is adjusted by the optimizer so that global minimum is not missed or overshooted.Rather than adjusting the variable learning rates based on average first moment (mean) as in RMSProp, Adam also uses the average second gradient moments with variance as well. The algorithm specifically calculates an exponential moving average of gradient , and the beta parameters control the decay rates of these moving averages. The original value of the moving averages and values of beta1 and beta2 close to 1.0 (recommended) results in a bias of estimates of the moment to zero. This bias is resolved by measuring biased calculations first before measuring bias-corrected forecasts afterwards.

The extracted neurons are then stacked to a MLP with 10 neurons in the hidden layer having an RELU activation function and a sigmoid function in the output layer.

Sample Size	Accuracy	AUC	F1 Score	Recall	Precision
20%	0.719	0.720	0.718	0.719	0.723
40%	0.721	0.721	0.721	0.721	0.721
60%	0.728	0.728	0.728	0.728	0.728
80%	0.736	0.736	0.735	0.736	0.739
100%	0.704	0.704	0.701	0.704	0.712

TABLE III
RESULTS OF STACKED AUTO ENCODER FOR EVERY BATCH SIZE FOR CENSUS DATA

An increasing pattern in accuracy could be observed from 20% to 80% batch sizes. However, the accuracy dropped when training with the 100% data.The results could be evaluated

with many inferences. Neural networks are algorithm with can extract non linear relationships, which means that the input predictors and output data points can learn complicated nonlinear relationships. They can accurately measure nonlinear functions that are challenging. Neural network designs are therefore referred to as having a low bias and high variance. They get a low bias due to the strategy making few generalizations about the mapping function's mathematical conceptual model. They have a high variance due to their sensitivity to examples used to train the model. Disparities in the examples of training can imply a very distinct consequent system with different abilities in turn. However, this difference in the behaviour of neural network on the same data is not due ti high-bias and low-variance but because the learning algorithm is stochastic.Examples of randomness include the small random values used before each training epoch to initialize the weights of the model and to randomly sort examples in the training data set. Instances of complexity include the random values used before each training epoch to initialize the weights of the model and to randomly sort examples in the training data set. This is useful randomness as it allows the model to "discover" good mapping function solutions automatically. Another possible rationale is that, small dataset results in low accuracy where the model might over fit the data and might not be the exact representation of the data. Large dataset could result in good accuracy however, the model might not have capacity to learn all the noises in the data or there is possibility that the dataset could have noises.

Model	Precision	Recall	F1 Score
Random Forest	0.83	0.85	0.84
Decision Tree	0.80	0.81	0.80
K-Nearest Neighbours	0.76	0.86	0.81

TABLE IV
RESULTS OF OTHER ALGORITHMS ON CENSUS INCOME DATASET.

For the same dataset other modern algorithms have been implemented to extract its maximum efficiency. It could be observed that Random Forest classifier performs better than other algorithms. A significant difference in the accuracy could be observed from stacked autoencoder with MLP model with the Random forest classifier. It is estimated that the model has lost some of its information due to the reduction in dimension of Autoencoder.

From the ROC curve, it could be observed that maximum area under the curve score was obtained when 80% of batch data was used and a steady decrease in the AUC score is seen when using 100% of the data. Even though, there are not significant difference between all of the batch sizes the maximum Accuracy was obtained by Random Forest algorithm.

c) Santander Customer Satisfaction data:

Santander customer satisfaction data is the largest dataset when compared with the other two datasets with 76020 observations and 370 features. Since the number of features are

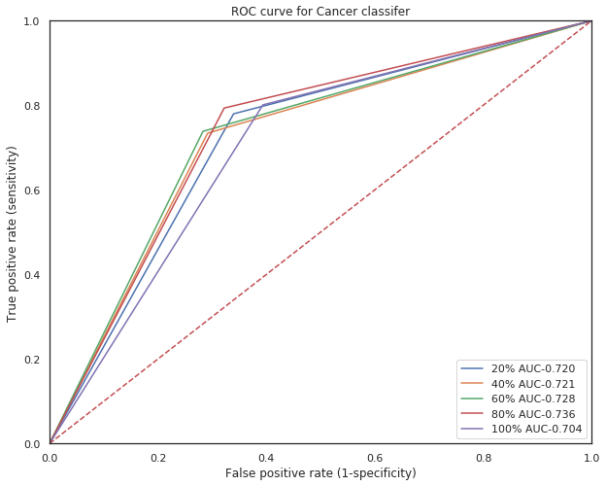

Fig. 7. ROC curve for Census data.

more than 100, it would be quite computationally expensive to train Autoencoder and encoder Stacked MLP models. A deep structure also has to be constructed for the model to capture all the patterns and useful informations.Hence, an Autoencoder of 370,300,200,150,200,300,370 has been chosen by comparing the validation loss and training loss with other structures.The count of bottle neck neurons that capture all the patterns is 150. Hence a dimensionality reduction is done from 370 to 150. Similar activation functions were used RELU for hidden layer and Sigmoid for output layer. However, the batch size has been increased to 128 for the algorithm to compute the data soon and converge faster. Epochs have been increased to 20 for both autoencoder and stacked model as epoch with 10 results in a very low AUC score. The stacked model has 3 hidden layers have with 30 neurons in each layer.

Sample Size	Accuracy	AUC	F1 Score	Recall	Precision
20%	0.69	0.69	0.69	0.69	0.69
40%	0.69	0.69	0.68	0.69	0.70
60%	0.70	0.70	0.69	0.70	0.70
80%	0.69	0.69	0.69	0.69	0.70
100%	0.73	0.73	0.73	0.73	0.73

TABLE V
RESULTS OF STACKED AUTO ENCODER FOR EVERY BATCH SIZE IN SANTANDER DATA

An intermittent pattern could be observed in the model accuracy. It was steadily increasing with a sharp decrease and again an major increase. The accuracy denotes the model performance for each batch size and the maximum accuracy of 73% were obtained at 100% data which endorses previous studies stating accuracy increased with larger sample sizes [27]. Highest precision and recall were also obtained when passing the entire dataset.

Model	Accuracy	Precision	Recall	F1 Score
XGBoost Classifier	89.05	0.89	0.89	0.89

TABLE VI
RESULTS OF OTHER ALGORITHMS ON SANTANDER DATASET.

The model has been compared with a XGboost classifier which achieved an accuracy of 89% in the test sample and the model was also used in the first place solution in Kaggle for this data.Unlike neural network XGboost gives more weights to the weak learners thereby constantly increasing the accuracy at each iteration. Neural network computes cost and updates the overall weight in the direction of the gradient.Due to computational inefficiency the stacked autoencoder model has not been trained with more than 20 epochs. However, increasing the number of layers and find tuning the hyper parameters could increase the accuracy of our stacked model.

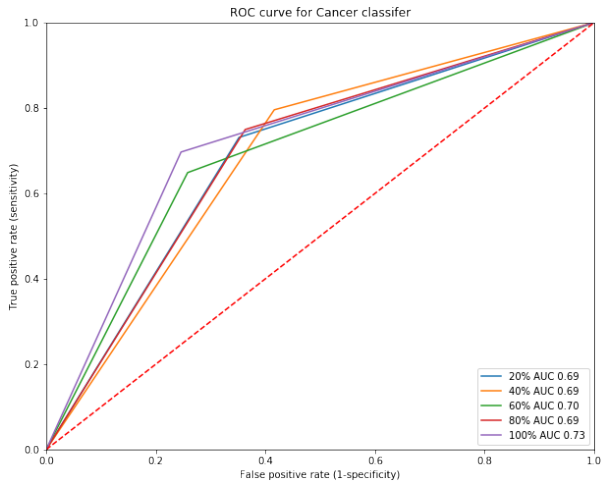

Fig. 8. ROC curve for Santander data.

The ROC curve for Santander data could be seen in the above chart, the maximum AUC score was 0.73 which was obtained when passing the full dataset. A constant and a steady increase of the AUC score is observed when the data sample size increases with intermittent drop in 80% of sample size.

VI. DISCUSSION

All the three datasets have imbalanced dataset so measuring the accuracy alone would not help us to access the true performance of the models. Metrics such as accuracy, accuracy, recall, F1 and Gini coefficient actully reveals the overall performance of the models. Our results show that the size of the data set affects the accuracy of stacked autoencoder models with neural MLP networks.In Breast cancer data, we were able to achieve the state of the art model's accuracy even after reducing its dimension to half of its original features. It also reduces the time for training when compared with training the network with all features. Our study also endorses the research [28].

Some of the limitations to be noted are that on random examples drawn from the data for each dataset, we do not report on the average accuracy of the model chosen. The algorithm and examples are fixed, and the learning algorithm is the only source of variation. The study shows how sensitive the model's approximately accuracy is to the test dataset size. This is an important factor as sometimes not enough attention is given to the size of the test set, using a familiar split for train / test splits of 70 percent/30 percent or 80 percent/20 percent.

Imbalanced sets of data are a common issue. While there are some methods like SMOTE that can create new data to equalize the data set, this data generation method is heavily dependent on sample distribution. When the sample distribution is very sparse, it is likely that the new data will differ from the space where the raw data is exited. Developing a technique for finding data mapping spaces based on the existing data distribution is crucial for generating information to equalize the data set, such as the currently popular deep generation model. Also an significant research path is the development of new algorithms to train unbalanced data sets. Secondly, the fluctuating accuracy can also be due to the increasing noise in the data for each batch.

Generally, having more training data will help to reduce the model error and variance leading to robust model without overfitting,but, in abstract, more data wont help if the model has high bias and it hence adding more data would not help.

Most obvious reason for the model accuracy falls for each batch is that variance in the model has exceeded its model capacity in terms of its layers and neurons defined. As more data being added to the neural network the performance of the neural network can be improved, but the model capacity have to be adjusted to endorse the data increment for each batch. There will eventually be a point of decreasing yields where more data will not provide more understanding on how to best model the issue of mapping. This point will be attained sooner than those in complex issues, such as categorizing photos of objects, for simpler problems, such as census data.

In future, a more robust deep learning frameworks could be tested for the same datasets and same objective. The number of hidden layers and neurons can be tuned completely for autoencoder and Stacked model using hyper parameter tuning. Due to lower computational configurations many of the experiments are not performed on the santander dataset. Different activation functions can also be tested in the models. Epochs could be increased to check till the model has been completely converged and optimal epoch could be found.Individual model structures were not tested for each batch size which could be developed in future to train model with its optimal parameters for each sample size. Also most of the algorithms which were compared with the stacked model was taken from previous research studies and public kernel platforms to access the latest models implied on the dataset. However, in future more advanced models can be implemented and its performance could be compared with the stacked model.

This work has underlined a key aspect of applied machine learning, namely that you need sufficient examples of both the issue to learn a helpful estimate of the unidentified mapping function.

VII. CONCLUSION

In this paper, a demonstration has been made on how autoencoder can be used as an dimensionality reduction technique and how the learned features can be used to used to predict the target classes. Also, the paper also discusses the impact of sample sizes on training a neural network by comparing it with a standard algorithms.The study also supports some of the past researches like increasing sample size increases the model accuracy and increases the variance in the model. Especially in census data, the model performance was decreased when training with whole dataset as confirms past studies where a point will be achieved where more data addition will not help the model. However, from the theoretical analysis aspect, such as dimensionality reduction , stackingand robustness, the proposed framework should be further investigated. Second, while the resultsin this paper shows some promising results, large-scale experiments, more complex data sets and statistical tests are needed to fully evaluate new framework capability. Third, in this paper, our discussion is limited to test sample sizes on stacked autoencoder model only and other neural networks were not tested. There are several new structures and techniques, such as cognitive neural network, and incremental learning, that have been established and applied in dimension reduction. It would therefore be very exciting to see how proposed structure works when integrated into modern systems and techniques

VIII. CODE REPOSITORY

I have placed the codes in the github repository mentioned below https://github.com/vincentkr18/ce888labs/tree/master/Project_1

IX. ACKNOWLEDGMENTS

Breast cancer, Census income and Santander Customer satisfaction dataset are available in Kaggle. Some of the state of the art techniques and approaches from other Kagglers to model and analyze the data were used and helpful in this research study. I thank Kaggle,Kagglers and other open source research communities for knowledge sharing.

REFERENCES

[1] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio., *Contractive auto-encoders: Explicit invariance during feature extraction. International Conference on Machine Learning*, 2011. 2, 3

[2] D. Rumelhart, G. Hinton, and R. Williams. *Learning internal representations by error propagation. Parallel Distributed Processing.* Vol 1: Foundations. MIT Press, Cambridge, MA, 1986. 2

[3] H. Lee, C. Ekanadham, and A. Ng. *Sparse deep belief net model for visual area v2. Advances in Neural Information Processing Systems*, 2008. 2

[4] G.E. Hinton and R.R. Salakhutdinov. *Reducing the dimensionality of data with neural networks.* Science, 313(5786):504, 2006.

[5] Chen, W.; Zheng, R.; Baade, P.D.; Zhang, S.; Zeng, H.; Bray, F.; Jemal, A.; Yu, X.Q.; He, J. *Cancer statistics in China, 2015*. CA Cancer J. Clin. 2016, 66, 115132. [CrossRef] [PubMed]

[6] Torre, L.A.; Bray, F.; Siegel, R.L.; Ferlay, J.; Lortet-Tieulent, J.; Jemal, A. *Global cancer statistics,* 2012. CA Cancer J. Clin. 2015, 65, 87108. [CrossRef] [PubMed]

[7] W. H. Cantrell, and W. A. Davis, "Amplitude modulator utilizing a high-Q class-E DC-DC converter", *2003 IEEE MTT-S Int. Microwave Symp. Dig.*, vol. 3, pp. 1721-1724, June 2003.

[8] Kaura, V. (2013), *Antecedents of Customer Satisfaction: A Study of Indian Public and Private Sector Banks, International Journal of Bank Marketing,* Vol. 31, No. 3, pp.167186.

[9] Ibok, I.N., and John, A.S. (2013), *Investigating Customer Satisfaction Driven Values in the Retail Banking Industry, International Journal of Finance and Accounting,* Vol. 2, No. 5, pp. 292-296.

[10] Singh, J., and Kaur, G. (2011), *Customer Satisfaction and Universal Banks: An Empirical Study, International Journal of Commerce and Management,* Vol. 21, No. 4, pp. 327-348.

[11] Fatima, J. and Razzaque, M.A. (2010a), *Service Quality, Customer Involvement and Customer Satisfaction: A Case Study of Retail Banking in Bangladesh,* Journal of Business and Policy Research, Vol. 7, No. 2, pp. 135146.

[12] Cui, Q. A., Wang, X., Li, H. J., and Kang, X. (2011), *Using PCA and ANN to Identify Significant Factors and Modeling Customer Satisfaction for Complex Service Processes*, in International Conference on Industrial Engineering and Engineering Management, China, 2011, IEEE, pp. 1800-1804.

[13] Goode, M.M.H., Davies, F., Moutinho, L., and Jamal, A. (2005), *Determining Customer Satisfaction from Mobile Phones: A Neural Network Approach,* Journal of Marketing Management, Vol. 21, No. 7-8, pp. 755-778.

[14] W. Wang, Y. Huang, Y. Wang, and L. Wang, *Generalized autoencoder: a neural network framework for dimensionality reduction,* in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,2014, pp. 490497.

[15] Z. Camlica, H. Tizhoosh, and F. Khalvati, *Autoencoding the retrieval relevance of medical images,* in Image Processing Theory, Tools and Applications (IPTA), 2015 International Conference on. IEEE, 2015, pp. 550555

[16] L. Meng, S. Ding, and Y. Xue, *Research on denoising sparse autoencoder,* International Journal of Machine Learning and Cybernetics, pp. 111, 2016.

[17] G. E. Hinton and R. R. Salakhutdinov, *Reducing the dimensionality of data with neural networks,* Science, vol. 313, no. 5786, pp. 504507, 2006.

[18] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, *An empirical evaluation of deep architectures on problems with many factors of variation,* in Proceedings of the 24th international conference on Machine learning. ACM, 2007, pp. 473480.

[19] I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, and A. Y. Ng,*Measuring invariances in deep networks, in Advances in neural information processing systems*, 2009, pp. 646654.

[20] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.- A. Manzagol, *Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, The Journal of Machine Learning Research,* vol. 11, pp. 33713408, 2010.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets, in Advances in Neural Information Processing Systems*, 2014, pp. 26722680.

[22] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, *Contractive auto-encoders: Explicit invariance during feature extraction,* in Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 833840.

[23] Z. Camlica, H. Tizhoosh, and F. Khalvati, *Autoencoding the retrieval relevance of medical images, in Image Processing Theory,* Tools and Applications (IPTA), 2015 International Conference on. IEEE, 2015, pp. 550555.

[24] H.-C. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach, *Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data,* IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8,pp. 19301943, 2013.

[25] Navoneel Chakrabarty, Sanket Biswas, *A Statistical Approach to Adult Census Income Level Prediction*

[26] William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. *Breast cancer Wisconsin (diagnostic) data set.* UCI Machine Learning Repository [http://archive. ics. uci. edu/ml/] (1992).

[27] "The effect of sample size and species characteristics on performance of different species distribution modeling methods Pilar A. Hernandez Catherine H. Graham Lawrence L. Master Deborah L. Albert"

[28] T. Fawcett. Roc graphs: Notes and practical considerations for researchers, 2004.

[29] "A Folded Neural Network Autoencoder for Dimensionality Reduction-Jing Wanga, Haibo Hea,, Danil V. Prokhorovb"

[30] "A Folded Neural Network Autoencoder for Dimensionality Reduction- Jing Wanga, Haibo Hea,, Danil V. Prokhorovb"
[31] "A Folded Neural Network Autoencoder for Dimensionality Reduction- Jing Wanga, Haibo Hea,, Danil V. Prokhorovb"