Auxiliary feature learning for small dataset regularization

1802525 - dy18708@essex.ac.uk

***Abstract*— The main objective of this research is to perform a comparative study between performance of discriminative neural network which are trained using selected auxiliary features from an autoencoder and performance of trained state of art models by augmented sampling of data.**

I. INTRODUCTION

With the development of the big data technology, we have been given more and more high-dimensional data, which really boost the performance of machine learning models. However, there are considerable noisy and useless features often collected or generated by different sensors and methods, which also occupy a lot of computational resources. Therefore, feature selection acts a crucial role in the framework of machine learning which removes nonsense features and preserves a small subset of features to reduce computational complexity.

In order to explore a more effective feature selection method, this paper propose to use an autoencoder network for selecting features with high representability, which is a widely used neural network for supervised dimensionality reduction [12]. Since the redundant features can be represented by linear or nonlinear combinations of other useful features, the autoencoder network can squeeze input features into a low-dimensional space and represent original features by exploiting these low-dimensional data. Therefore, features with less effect on the low-dimensional data (i.e., hidden units) could be recognized as redundancy, which can be removed by a group sparsity regularization. Experiments conducted on benchmark datasets verify the effectiveness of the proposed method over other methods

II. BACKGROUND

This section provides a brief review of past work on data augmentation and Auxiliary features.

A. Data Augmentation

The problem with small datasets is that models trained with them do not generalize well data from the validation and test set. Hence, these models suffer from the problem of overfitting. The reduce overfitting, several methods have been proposed [13].

Data augmentation is one of the way we can reduce overfitting on models, where we increase the amount of training data using information only in our training data. The field of data augmentation is not new, and in fact, various data augmentation techniques have been applied to specific problems. The main techniques fall under the category of data warping, which is an approach which seeks to directly augment the input data to the model in data space. The idea can be traced back to augmentation performed on the MNIST set in [14].

B. Auxiliary feature learning

In terms of auxiliary feature learning this paper work is related to autoencoders.Valpola (2015)[1] presents the Ladder Network model that allows efficient training of deep autoencoders by eliminating the requirement that deeper layers retain low-level representations. This is done by having two identical encoder networks, with the exception that noise is added to each layer of activation. A denoising decoder tries to reconstruct the noisy encoder's clean input, just like a denoising autoencoder would. Lateral links between the corresponding layers in the noisy network and the decoder network allow the decoder to have access to low-level representations, eliminating the need for deeper encoder layers to retain this information. In addition, each layer in the networks is separately trained. The clean encoder network layers serve as targets for the decoder network layers and vice versa. Finally, the clean encoder target output is used as the noisy encoder target output.

(Rasmus, Valpola, Honkala, Berglund, and Raiko 2015) [2]build on Valpola's proposed Ladder Network and demonstrate that it can be used for semi-supervised learning. Their network can learn simultaneously in both supervised and unattended settings-enabling supervised learning to provide the context for the "right" representations that unattended learning should learn. The MNIST, permutation invariant MNIST and CIFAR10 datasets were tested on their network. An impressive result was the achievement of a error rate of 1.06 percent on the MNIST dataset using only 100 labeled training examples.

Raina et al. introduced the paradigm of "self learning" [3], in which a large number of unlabeled images are downloaded from the World Wide Web in order to learn good feature representations and improve performance in a given task of classification of computer vision.

III. METHODOLOGY

The methodology involves the data collection from online database. The description of the data and possible relationships between variables would be investigated under Data description and Data exploration sections. Essential pre-processing steps such as missing value imputations, categorical encoding, Normalization will be carried out in Data preprocessing section. Modelling section deals with principles of models used to support this research objective. Finally, Evaluation section to discuss about performance evaluation metrics.

A. Data Collection

Data scientists need large amounts of data to apply and develop their new research ideas. However, valuable business data are not freely available most of the time, so that a data expert cannot always have access to real data. Competition is usually an opportunity for data miners to access real business data and compete with others to find the best data technique. The Kaggle website (https: / www.kaggle.com/) is a web platform for companies to post and review their data by data scientists. This allows data experts to access real data sets and solve problems with the possibility of winning a company award.

The dataset I referred as Life Insurance risk classification Dataset consist of life insurance claims from the Prudential Insurance Company, and were posted for the Kaggle competition called the "prudential life insurance Challenge", The contests goal was to develop a simplified model for quickly and accurately binning life insurance applicants into risk classes or profiles.

The dataset II referred as Driver Insurance Claim Dataset consist of driver insurance claims from the Porto Seguro Insurance Company, and were posted for the Kaggle competition called the "Predicting Insurance Claims in Brazil", The contests goal was to predict whether the customer will le an insurance claim during a period of interest.

The dataset III referred as Motor insurance claim prediction consist of motor insurance claims , and were posted for the Kaggle competition called the "Motor Insurance", The contests goal was to predict whether the customer will le an motor insurance claim during a period of interest.

B. Data Description

1) Life Insurance Dataset: The data set consists of 59,381 applications with 128 attributes describing the characteristics of applicants for life insurance. The data set includes anonymized nominal, continuous and discrete variables. Table I describes the data set variables.

2) Driver Insurance Claim Dataset: The training dataset contains 595,213 customer records. Each record consists of 57 unknown features and a target indicating whether the customer has submitted a claim. 21,694 examples have label 1, while the other 573,518 have label 0. The test set contains 892,816 label-free records. Of the 57 features available to each customer, 26 represent permanent or ordinal values and 31 represent categorical values.This project faces the unique challenge of fitting a dataset with unlabeled features. The unique challenge of fitting a dataset with unlabeled features is facing this project. All features are unlabeled in order to protect the identities of car insurance holders in the data provided by Porto Seguro, one of the largest car and homeowner insurance companies in Brazil. Some data type information is provided, however, and features are grouped by type. Since no additional information on the features has been provided, we can not rely on feature selection intuition. Table II describes the data set variables.

Attributes	Type	Description
Id	Numeric	A unique Identifier associated with an application
Product Info 1-7	Categorical	7 A set of normalized attributes concerning the product applied for
Ins Age	Numeric	Normalized age of an applicant
Ht	Numeric	Normalized height of an applicant
Wt	Numeric	Normalized weight of an applicant
BMI	Numeric	Normalized Body Mass Index of an applicant
Employment Info 1-6	Numeric	A set of normalized attributes concerning employment history of an applicant
Insured Info 1-6	Numeric	A set of normalized attributes offering information about an applicant
Insurance History 1-9	Numeric	A set of normalized attributes relating to the insurance history of an applicant
Family Hist 1-5	Numeric	A set of normalized attributes related to an applicants family history
Medical History 1-41	Numeric	A set of normalized variables providing information on an applicants medical history
Medical Keyword 1-48	Numeric	A set of dummy variables relating to the presence or absence of a medical keyword associated with the application
Response	Categorical	This is a target variable, which is an ordinal measure of risk level, having 8 levels

TABLE I
LIFE INSURANCE DATA DESCRIPTION

3) Motor insurance Claim prediction Dataset: The data set consists of 40,000 transaction records with 101 attributes. All predictor variables are unlabelled which are represented as x0,x1,x2 till x100. and response variable as y.

C. Data Exploration

As the Qualitative inputs gives the qualitative outputs, Data Exploration has become a crucial step in obtaining information of those inputs by exploring missing values, data outliers and categorical variables.

1) Missing values: The first thing is to check if there is any missing value in the dataset. Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we cannot analyse the behaviour and relationship with other variables correctly. It can lead to wrong prediction or classification. These missing values can be occurred during data extraction or data collection.

Below tables contains the count as well as the percentage of missing values.

Table III gives information on missing values in life insurance dataset .As displayed, there are few variables whose missing values are at 98 percent, which represents data have a severe problem of missing information.

Table IV gives information on missing values of driver insurance claim dataset .As shown

This dataset of Motor insurance Claim prediction contains missing values in every variable which is not possible to

Attributes	Type	Description
Id	Numeric	A unique Identifier associated with an application
Target	Categorical	7 normalized attributes concerning the product applied for
Ps ind 01-18	Numeric	Normalized age of an applicant
Ps ind 01-18 cat	Categorical	Normalized height of an applicant
Ps ind 01-18 bin	Binary	Normalized weight of an applicant
Ps reg 01-03	Numeric	Normalized Body Mass Index of an applicant
Ps reg 01-03 cat	categorical	6 normalized attributes concerning employment history of an applicant
Ps reg 01-03 bin	binary	6 normalized attributes offering information about an applicant
Ps car 01-15	Numeric	9 normalized attributes relating to the insurance history of an applicant
Ps car 01-11 cat	categorical	5 normalized attributes related to an applicants family history
Ps calc 01-15	Numeric	41 normalized variables providing information on an applicants medical history
Ps calc 01-11 bin	binary	48 dummy variables relating to the presence or absence of a medical keyword associated with the application

TABLE II

DRIVER INSURANCE CLAIM DATA DESCRIPTION

Features	Missing Count	missing count percentage
Employment Info 1	19	0.031997
Employment Info 4	6779	11.416110
Employment Info 6	10854	18.278574
Insurance History 5	25396	42.767889
Family Hist 2	28656	48.257860
Family Hist 3	34241	57.663226
Family Hist 4	19184	32.306630
Family Hist 5	41811	70.411411
Medical History 1	8889	14.969435
Medical History 10	58824	99.061990
Medical History 15	44596	75.101463
Medical History 24	55580	93.598963
Medical History 32	58274	98.135767

TABLE III

MISSING VALUES IN LIFE INSURANCE DATASET DATA

features	Missing Count	missing count percentage
ps ind 04 cat	19	0.139775
ps ind 05 cat	6779	9.782590
ps reg 03	10854	181.492397
ps car 01 cat	25396	0.180192
ps car 02 cat	28656	0.008420
ps car 03 cat	34241	692.529597
ps car 05 cat	19184	448.882639
ps car 07 cat	41811	19.347940
ps car 09 cat	8889	0.958219
ps car 11	58824	0.008420
ps car 12	44596	0.001684
ps car 14	55580	71.773800
ps ind 02 cat	216	0.363753

TABLE IV

MISSING VALUES IN DRIVER INSURANCE CLAIM DATASET DATA

display in tabular form, but they are of very less percentage ranging from 0.0275 percentage to 0.035 percentage.

2) Outlier Detection: Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample. Outliers can drastically change the results of the data analysis and statistical modelling. There are numerous unfavourable impacts of outliers in the data set includes increases the error variance and reduces the power of statistical tests. Most commonly used method to detect outliers is visualization. We use various visualization methods, like Box-plot, Histogram, Scatter Plot (above, we have used box plot and scatter plot for visualization). Most of the ways to deal with outliers are similar to the methods of missing values like deleting observations, transforming them, binning them, treat them as a separate group, imputing values and other statistical methods.

3) Categorical variables: Variables containing label values are categorical data rather than numerical values. Categorical features are common in many domains and are often highly cardinal. Many algorithms for machine learning cannot operate directly on label data. All input variables and output variables are required to be numerical. In general, this is mostly a constraint on the algorithms themselves to effectively implement machine learning algorithms rather than tough limitations. This means that it is necessary to convert categorical data into a numerical form. If the categorical variable is a variable of output, you may also want to convert the model's predictions back into a categorical form to present or use them in some application. Under such circumstances, the use of one-hot encoding results in very high dimensional representations of vectors, causing concerns about memory and computability for machine learning models.

D. Data Pre-processing

Upon Data exploration it is evident that all there datasets have considerable amount of missing values and categorical variables. The data will be cleaned in this step in order to make the data consistent with the analysis.

1) Missing values: There are several ways that missing values can be calculated such as case analysis, imputation and missing indicator.

Because of its large percentage of missing values in Prudential Life Insurance dataset, columns named Medical History 10, Medical History 24 and Medical History 32 are just dropped from the data. For the rest of the columns mean and mode imputation techniques are used for numerical and categorical variables respectively.

Driver Insurance Claim Dataset contains numerical and categorical variables, which are handled using mean and mode imputation techniques.

Finally, variables in Motor insurance Claim prediction Dataset contains very less percentage of missing values. Henceforth, dropping all missing values can make the data to be consistent.

2) Categorical variables: This process includes selection of all categorical variables and one-hot coding has been applied to that variable in every datasets , as a result all categorical variables have been converted to numerical form.

One-hot encoding is the most common approach to converting categorical features into an appropriate format for use as input to a model of machine learning. An interesting feature of the one-hot encoding is that the categories are represented as independent concepts-one way to see this is to note that the internal product between any two vectors is zero, each vector in Euclidean space is equally distant from each other. Since data using one-hot encoding is numerical in nature, such categorical feature information can easily be incorporated by a machine learning model by learning a separate parameter, w, for each dimension.

3) Data Normalization: Standardization is a scaling technique or a pre-processing mapping technique. Where, from an existing one range, we can find new range. It can be very helpful for the purpose of predicting or forecasting. As we know, there are so many ways to predict or predict, but they can all vary a lot. Therefore, it is necessary to keep the large variation of prediction and to forecast the technique of normalization to make them closer. Before modelling, data is now standardized. In order to make the dataset well-structured or structured, we have proposed one technique called Min-Mix Normalization that gives one dataset scaled or transformed or structured or standardized for our research work within range 0 and 1. This is done by subtracting the sample's minimum value and dividing it with the sample range.

E. Models

Initially, an auto encoder has been trained with small sample of dataset. Next, a discriminative neural network has been trained with the features extracted from auto encoder. Finally, a few State Of Art models are trained with the same data supplied to neural network

1) Auto Encoders: The basic framework of autoencoder [Bengio, 2009] is a feed forward neural network with an input layer, an output layer and one or more hidden layers between them. An autoencoder framework usually includes the encoding and decoding processes. Given an input x, autoencoder first encodes it to one or more hidden layers through several encoding processes, then decodes the hidden layers to obtain an output x. Autoencoder tries to minimize the deviation of x from the input x.Fig 1 shows a naive example of the architecture of the autoencoder.

The data features are entered into the network's input and output layer. Therefore, when the autoencoder is trained, the hidden layer in theory contains all information with fewer nodes inputs. After training the autoencoder, we keep the autoencoder's weights, biases and nodes and connect the hidden layer as inputs to the main network . This pre-training process reduces the size of the main network inputs, leading to higher efficiency and better performance, as it reduces the calculation of the main network requirement and the noise of the original inputs.

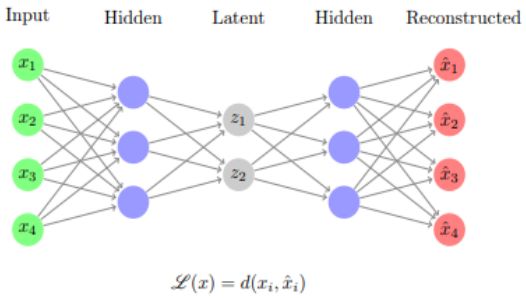

Fig. 1. Autoencoder

2) Neural Network: Neural network is an algorithm that simulates the neural system of animals. It has nodes (hidden units) and neuron-and axon-related connections as shown in the following figure. Nodes are places where data (number) can be temporarily stored and the connections are weights and biases. When the inputs pass through the connections as a flow, the weights are multiplied and the biases added, and the sum of these results is added and stored in the next node. When the data flows to the output layer, the actual result is compared. The difference or so-called cost entropy is then calculated and the network attempts to adjust those weights and biases to reduce the cost. When the cost is small enough, the network is well trained.Fig 2 shows a naive example of the architecture of the Neural network.

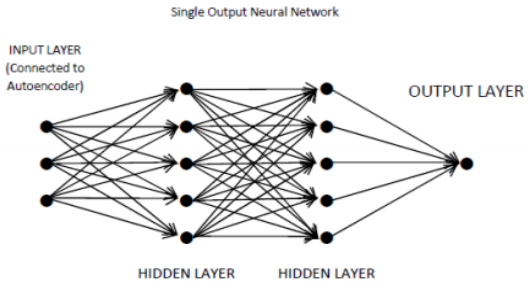

Fig. 2. Neural network

After the network is trained, we can test the network by predicting target variables given features as input.

3) State Of Art model: Random Forest is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees which results in more accurate and stable prediction.

Logistic regression is a method for classifying data into discrete outcomes. This approach allows an algorithm being used in a machine learning application to classify incoming data based on historical data

F. Performance Evaluation

The next step is to find out how effective is the model based on performance evaluation metrics. Choice of metrics

influences how the performance of machine learning algorithms is measured and compared. We can use classification performance derived from simple confusion matrix.Fig 3 shows a architecture of the confusion matrix.

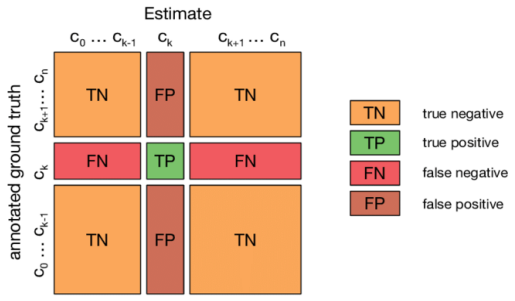

Fig. 3. Confusion matrix

1) Confusion matrix: The confusion matrix, also known as the contingency table, is a specific layout of the table fig 6 that shows the classification test performance. It contains rows and columns that report the number of false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN)

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{P+N} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \quad (1)$$

(1)It is the ratio of the number of correct predictions to the total number of input samples

$$Recall/Sensitivity = \frac{tp}{tp+fn} \quad (2)$$

(2)Out of all the positive classes, how much we predicted correctly are referred as recall

$$Precision = \frac{tp}{tp+fp} \quad (3)$$

(3)Out of all the classes, how much we predicted correctly are referred as precision.

$$F_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}} = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \quad (4)$$

(4)It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time.

Sensitivity and specificity are two components that measure the inherent validity of a test. Receiver operating characteristic (ROC) curve is the plot that depicts the trade-off between the sensitivity and (1-specificity) across a series of cut-off points.

IV. EXPERIMENTS

A sample of train data is used to train the Auto encoder-decoder network. The inputs are passed to encoder phase where the nodes in each layer are stacked in descending order. Further fed to decoder phase which is mirror image of encoding phase ,nodes in each layer are stacked in ascending order. The input signal and output of decoder are compared over various iterations to get a minimum reconstruction error. Once the network trained with minimum reconstruction error, the auxiliary features of sample data are collected from the encoding phase. These features are further used to train a discriminative neural network. This process is implemented with all three datasets on basis of data augmentation. A typical encoder structure is as follows:

		Dense Layers	Input features	Output features
Life Insurance Risk Classification	Encoder phase	1st layer	123	123
		2nd layer	123	64
		3rd Layer	64	32
		4th layer	32	16
Driver Insurance Claim Prediction	Encoder phase	1st layer	218	218
		2nd layer	218	64
		3rd Layer	64	32
		4th layer	32	16
Motor insurance Claim Prediction	Encoder phase	1st layer	129	129
		2nd layer	128	64
		3rd Layer	64	32
		4th layer	32	16

Fig. 4. AutoEncoderNetworkModel

The 1st layer of the encoder takes dataset predictor variables as input and passes to second dense layer. Similar approach has been carried forward till end of layers. Finally the auxiliary features of sample data are collected from the encoding phase. These learned features are further used to train a discriminative neural network. Similarly the data has been trained on a state of art model to perform a comparative study.

This training process has been performed starting with a small sample of data and gradually increasing to entire data. Once the training has been completed, performance has been network is identified on test data. Different evaluation metrics are captured at each sampling rate.

Figures from 4 to 8 shows comparitative study of evaluation metrics performance of discriminative neural network which are trained using selected auxiliary features from an autoencoder and performance of trained state of art models by augmented sampling of data

Sample rate	Neural Network trained with Auxilary Features				State Of Art Model(Random Forest)			
	Accuracy	Precision	Recall	F-score	Accuracy	Precision	Recall	F-score
10	0.33	0.11	0.33	0.16	0.41	0.39	0.41	0.39
20	0.38	0.20	0.38	0.26	0.43	0.41	0.43	0.41
30	0.36	0.17	0.36	0.23	0.43	0.42	0.43	0.42
40	0.37	0.21	0.37	0.25	0.43	0.41	0.43	0.42
50	0.39	0.32	0.39	0.30	0.44	0.42	0.44	0.42
60	0.41	0.29	0.41	0.32	0.43	0.41	0.43	0.41
70	0.41	0.34	0.41	0.32	0.43	0.42	0.43	0.42
80	0.41	0.34	0.41	0.33	0.42	0.40	0.42	0.41
90	0.40	0.32	0.40	0.31	0.45	0.43	0.45	0.43
99	0.41	0.33	0.41	0.32	0.44	0.42	0.44	0.42

Fig. 5. EvaluationComparision of NN and SOAM in Life Insurance

The state of art model performs better even the data is very less when compared with the neural network architectures. The Neural networks performance is increasing with increase in sampling rates.

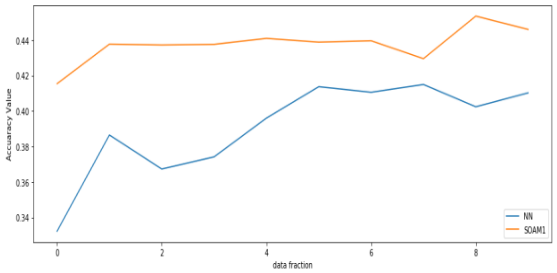

Fig. 6. Accuracy Comparision of of NN and SOAM in LifeInsuranceRiskClassification

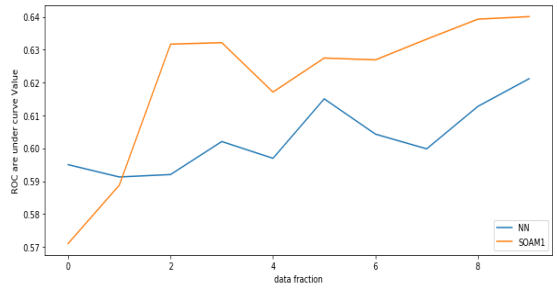

Fig. 7. ROC Comparision of of NN and SOAM DriverInsuranceClaim

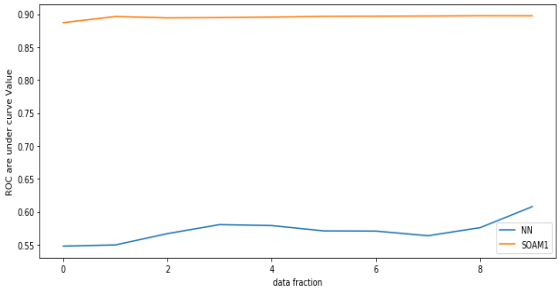

Fig. 8. ROC Comparision of of NN and SOAM MotorinsuranceClaim

V. DISCUSSIONS

Initially, an auto encoder has been trained with small sample of dataset such that reconstruction error is minimum. These learned features are further used to train a discriminative neural network. Similarly the data has been trained on a state of art model to perform a comparative study.For binary classification data sets, the output Sigmoid layer will be used at the end of the neural network and the Softmax layer will be used for multi-label classification.This task is performed on there datsets which are having different objectives to acheive. In addition, tuning the Auto-Encoder Network and Discriminative Classifier hyper parameters is challenging. The performance of the model depends heavily on the hyper parameters like number of auxiliary features, learning rate, activation function.

VI. CONCLUSIONS

The amin of this paper is to perform a comparative study between performance of discriminativeneural network which are trained using selected auxiliaryfeatures from an autoencoder and performance of trained stateof art models by augmented sampling of data.The three different kinds of data sets used to perform this task and it is evident that the state of art model performs better even the data is very less when compared with the neural network architectures. The Neural networks performance is increasing with increase in sampling rates.

REFERENCES

[1] Harri Valpola. From neural PCA to deep unsupervised learning. In Adv. in Independent Component Analysis and Learning Machines, pages 143171. Elsevier, 2015. arXiv:1411.7783.

[2] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. 2015. Semisupervised learning with Ladder networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15), C. Cortes, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). MIT Press, Cambridge, MA, USA, 3546-3554.

[3] R. Raina, A. Battle, B. Packer, H. Lee and A. Ng. Self-taught learning: Transfer learning from unlabeled data. Prc. ICML , 2007.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee and A. Ng. Multimodal Deep Learning Prc. ICML , 2011.

[5] Dan Huangfu. Data mining for car insurance claims prediction. 2015.

[6] Andrea Dal Pozzolo. Comparison of data mining techniques for insuranceclaim prediction. 2010.

[7] Cummins Phillips 1999 Lee Urrutia 1996.Analysis and Prediction ofInsolvency in the Property-Liability Insurance Industry: A Comparison ofLogit and Hazard Models. The Journal of Risk and Insurance,. 63(1), 121-130. doi:10.2307/253520.

[8] Noorhannah Boodhun and Manoj Jayabalan. Risk prediction in life insur-ance industry using supervised learning algorithms.Complex IntelligentSystems, 4(2):145154, 2018.

[9] Dasheng Gu, Jingwei Shen, and Xinyuan Wang. Deep learning.Instructor,2018

[10] Noorhannah Boodhun and Manoj Jayabalan. Risk prediction in life insur-ance industry using supervised learning algorithms.Complex IntelligentSystems, 4(2):145154, 2018.

[11] Dasheng Gu, Jingwei Shen, and Xinyuan Wang. Deep learning.Instructor,2018

[12] Shuyang Wang, Zhengming Ding, and Yun Fu, Feature selection guided auto-encoder., in AAAI, 2017, pp. 27252731

[13] B. Wang and D. Klabjan. Regularization for unsupervised deep neural nets. CoRR, abs/1608.04426, 2016

[14] Cummins Phillips 1999 Lee Urrutia 1996.Analysis and Prediction ofInsolvency in the Property-Liability Insurance Industry: A Comparison ofLogit and Hazard Models. The Journal of Risk and Insurance,. 63(1), 121-130. doi:10.2307/253520.

[15] Willis R. Brooks M Smith, K.An Analysis of Customer Retention andInsurance Claim Patterns Using Data Mining: A Case Study. The Journalof the Operational Research Society,The Journal of the Operational ResearchSociety,. 51(5), 532-541. doi:10.2307/254184

[16] Smith K. A. Willis R. J. Yeo, A. C. and M Brooks.Clustering technique forrisk classication and prediction of claim costs in the automobile insurance in-dustry. Int. J. Intell. Syst. Acc. Fin. Mgmt.,. 10: 3950. doi:10.1002/isaf.196

[17] Cedric Seger. An investigation of categorical variable encoding techniquesin machine learning: binary versus one-hot and feature hashing, 2018

[18] Gavin Brown. Ensemble learning.Encyclopedia of Machine Learning, pages312320, 2010.

[19] Gareth James.Majority vote classifiers: theory and applications. PhD thesis,Stanford University, 1998.

[20] Iwan Syarif, Adam Prugel-Bennett, and Gary Wills. Svm parameter opti-mization using grid search and genetic algorithm to improve classificationperformance.Telkomnika, 14(4):1502, 2016.

[21] Marc Claesen and Bart De Moor. Hyperparameter search in machine learn-ing.arXiv preprint arXiv:1502.02127, 2015.

[22] Dan Huangfu. Data mining for car insurance claims prediction. 2015.

[23] Andrea Dal Pozzolo. Comparison of data mining techniques for insuranceclaim prediction. 2010

[24] Matthew Millican, Laura Zhang, and Dixee Kimball. Cs 229 final report:Predicting insurance claims in brazil. 2017.

[25] H. S. Baird. Document image analysis. chapter Document Image Defect Models, pages 315325. IEEE Computer Society Press, Los Alamitos, CA, USA, 1995.

[26] Saba Arslan Shah and Mehreen Saeed. Predicting purchased policy for cus-tomers in allstate purchase prediction challenge on kaggle.

[27] Essam Shaaban, Yehia Helmy, Ayman Khder, and Mona Nasr. A proposedchurn prediction model. 2012.

[28] U. Sivarajah, M. Kamal, Z. Irani, and V. Weerakkody, Critical analysis of Big Data challenges and analytical methods, Journal of Business Research, vol. 70, pp. 263-286, Aug. 2017.

[29] Y. Joly, H. Burton, Z. Irani, B. Knoppers, I. Feze, T. Dent, N. Pashayan, S. Chowdhury, W. Foulkes, A. Hall, P. Hamet, N. Kirwan, A. Macdonald, J. Simard, and I. Hoyweghen, Life Insurance: genomic-sStratification and risk classification, European Journal of Human Genetics, vol. 22, pp. 575-579, May. 2014

[30] K. Umamaheswari, and D. Janakiraman, Role of data mining in Insurance Industry, An International Journal of Advanced Computer Technology, vol. 3, pp. 961-966, 2014

[31] A. Raj, and P. Joshi. (2017) Changing face of the Insurance Industry. [Online]. Available: https://www.infosys.com/industries/insurance/white-papers/Documents/changing-face-insurance-industry.pdf.

[32] J. Cummins, B. Smith, R. Vance, and J. Vanderhel, Risk classificaition in Life Insurance, 1st ed, New York: Springer-Science and Business Media, 2013.

[33] A. Bhalla, Enhancement in predictive model for insurance under-writing, International Journal of Computer Science Engineering Technology, vol. 3, pp. 160-165, May. 2012.

[34] K. Mishr, Fundamentals of life insurance theories and applications, 2nd ed, Delhi: PHI Learning Pvt Ltd, 2016

[35] A. Wuppermann, Private information in life insurance, annuity and health insurance markets, The Scandinavian Journal of Economics, vol. 119, pp. 1-45, May. 2016.

[36] A. Prince, Tantamount to fraud?: Exploring non-disclosure of genetic information in life insurance applications as grounds for policy rescis-sion, Health Matrix, vol. 26, pp. 255-307, 2016.

[37] D.M.Z. Mamun, K. Ali, P. Bhuiyan, S. Khan, S. Hossain, M. Ibrahim, and K. Huda, Problems and prospects of insurance business in Bangladesh from the companies perspective, Insurance Journal of Bangladesh Insurance Academy, vol. 62, pp. 5-164, Apr. 2016.

[38] T. Harri, and A. Yelowitz, Is there adverse selection in the life insur-ance market? Evidence from a representative sample of purchasers, Economics Letters, vol. 124, p. 520-522, Sept. 2014.

[39] D. Hedengren, and T. Stratmann, Is there adverse selection in life insurance markets? Economic Inquiry, vol. 54, pp. 450-463, Jan. 2016.

[40] JFernandez-Villacanas J. Segovia-Vargas M. J. Bousono-Calzon C Sal-cedoSanz, J. S. Genetic programming for the prediction of insolvency in non-life insurance companies. Computers Operations Research.

[41] IBM Reserah Article on Recent Increase in Data: https://www.ibm.com/blogs/insights-onbusiness/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retailmanage-it/

[42] Credit Card Fraud Detection Paper:http://ijcttjournal.org/Volume4/issue-7/IJCTT-V4I7P143.pdf

[43] Jason Wang and Luis Perez. The effectiveness of data augmentation inimage classification using deep learning.Convolutional Neural Net-worksVis. Recognit, 2017.

[44] Kai Han, Yunhe Wang, Chao Zhang, Chao Li, and Chao Xu. Autoen-coder inspired unsupervised feature selection. In2018 IEEE InternationalConference on Acoustics, Speech and Signal Processing (ICASSP), pages29412945. IEEE, 2018.

[45] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and QingHe. Supervised representation learning: Transfer learning with deep au-toencoders. InTwenty-Fourth International Joint Conference on ArtificialIntelligence, 2015