Auxiliary feature learning for small dataset regularization

1802525 - dy18708@essex.ac.uk

***Abstract*— The overall objective of this project is to check whether one can utilize generic auxiliary tasks to learn features that help regularise the network. The proposed model is composed of 5 tasks which are; dataset selection and exploration, training an autoencoder, training standard discriminative neural network with learned features, comparing AUC scores and finally comparing results with state-of-art methods. The primary focus of this paper is to perform task of dataset selection and exploration. Three datasets from insurance domain has been chosen to accomplish this objective.**

I. INTRODUCTION

A key challenge for the insurance industry is to charge each customer an appropriate price for the risk they represent. Risk varies widely from customer to customer, and a deep understanding of different risk factors helps predict the likelihood and cost of insurance claims. Improving the accuracy of insurance claims benets both customers and insurance companies .The goal of these datasets is to see how well various statistical methods perform in predicting Insurance claim payments based on the characteristics of the customers vehicles, driving abilities and background details. The rest of the paper is organized as follows. Section II introduces the background and literature review of autoencoders and discriminative neural network .Section III details the methods used for data collection and data exploration.Section IV describes furthur work to be carried to accomplish objective .In section V, experiments shows acquired results of data exploration .Important discussions and conclusions are respectively presented in Section VI and Section VII.

II. BACKGROUND

Autoencoders are an unsupervised class of neural networks that seek to represent the input in a lower dimensional space before restoring it to its original shape. They consist of an encoder network encoding the input into the low-dimensional space, followed by a decoder network trying to reconstruct the original input from the low-dimensional display. These models can be slow to train, as deep layers in the encoder need to retain information about the low-level characteristics of the data, which means that high-level abstractions are difficult to learn. This is contrasted with supervised classification networks, in which later layers that learn higher-level representations can discard low-level representations in the early layers. Autoencoders also have a problem with learning useful representations, which can vary depending on the data domain. The denoted autoencoders finally add noise to the input and minimize the error between the reconstructed, noisy input and the original, denoted input.

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.

Although much of the work in unsupervised feature learning and deep learning has focused on learning features for single modalities, work has been done on multiple modalities. In particular, Ngiam et al. used unattended learning features and deep learning techniques to learn audio and video features[4]. They have shown an example of cross-modality learning in which better video features can be learned if audio features are present during the learning time of the feature. We use a similar approach to learn better features of computer vision by using text information during the learning time of the feature.

III. METHODOLOGY

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 can not 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 (http / 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 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 Claim amount Prediction Dataset consist of automobile insurance claims from the Allstate Insurance Company, and were posted for the Kaggle competition called the "Claim Prediction Challenge". The goal of the competition was to predict the amount of money the insurance has to pay to its clients.

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.

3) Claim amount Prediction Dataset: The data provided by the competition organizer consists of a training set for the presentation of actual claims and a test set for which the claims are not present and must be predicted. The training data set contains some missing values and has 2005-2007 information, while the test data has 2008 information. For insured vehicles, each row contains information worth one year. The objective of the competition is to improve the ability to accurately predict the payment of insurance claims using vehicle characteristics, which is the response variable (the dollar amount of claims experienced in that year for that vehicle). In the set of independent variables, non-vehicle characteristics are included and labeled as NV. These non-vehicle variables are not expected to make major contributions to the model, but interesting interactions can occur between these variables and the vehicle variables. The data set variables are displayed in the following table III

"CalendarYear" refers to the year the vehicle was insured. "Household ID" is a household identification number that enables each household to be tracked year - to-year. Since a customer can insure multiple vehicles in one household, each household identification number can be associated with multiple vehicles. "vehicle" identifies these 38 vehicles (the same number of vehicles may not apply from year to year to the same vehicle). The remaining columns contain various vehicle characteristics and other insurance policy characteristics. The data set consists of both continuous and categorical variables. Since the significance of column Var1, Var2, etc. Isn't given it is not clear their importance in the problem. Of all 32 variables, 16 are categorical in the dataset. The training set instances with a Claim variable greater than zero are only 0.73 percent, so the dataset is highly unequal. The response variable "ClaimAmount" is referred to in this work as Y and input variables are referred to as X. Predicting

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

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

Attributes	Type	Description
Id	Numeric	A unique Identifier associated with an application
HouseholdID	Numeric	Household identification number
Vehicle	categorical	Regisered vehicles
CalendarYear	Numeric	Year of registration
Model Year	Numeric	Year of lauch
Cat1-12	categorical	A set of normalized attributes offering information about an applicant
Var1 8	Numeric	A set of normalized attributes offering information about an applicant
NVCat	categorical	A set of normalized attributes offering information about an non-vehicle variables
NVVar1-4	Numeric	A set of normalized attributes offering information about an non-vehicle variables
ClaimAmount	Numeric	Tgis is the target variable,which is the measure of claim amount

TABLE III
CLAIM AMOUNT PREDICTION DATASET DATA DECRIPTION

the value of the claim can be seen as a problem of regression because the result is a continuous variable. At the same time, if its value is greater than zero and zero otherwise (Y= 0), the problem could be transformed into a classification by transforming the claim amount variable into a binary variable taking value one (Y= 1).

C. Data Exploration

Quality of the inputs gives the quality of outputs which made Data Exploration as a crucial step in solving any sort of problems. It gives us insights on missing values ,outlier detection ,correlation analysis which are essential steps in data-pre-processing to make inputs as qualitative.

1) Missing values: 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 behavior 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. Either Deletion or imputation techniques or prediction modelling techniques will be used to deal with such missing values.

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) Inbalanced target distribution check: Imbalanced data typically refers to a problem with classification problems where the classes are not represented equally.For example, we may have a binary classification problem with 100 observations.A total of 80 instances are labeled with Class-1 and the remaining 20 instances are labeled with Class-2. This is an imbalanced dataset and the ratio of Class-1 to Class-2 instances is 80:20 or more concisely 4:1.we can have a class imbalance problem on two-class classification problems as well as multi-class classification problems. Most techniques can be used on either.Collecting more data, changing performance metric, Resampling dataset , generating synthetic samples ,choosing penalized model are some of the solutions to overcome imbalance in datasets

IV. FURTHUR WORK

Before we feed these inputs into the network, we feed them into an autoencoder first. An autoencoder is a symmetric network of neural inputs and outputs. The input is compressed and noise reduced. Autoencoders are multi-layer perceptrons based on matrix factorization models. Fig 2 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 below. This pre-training process reduces the size of the main network inputs, leading to higher efficiency and better performance, as it

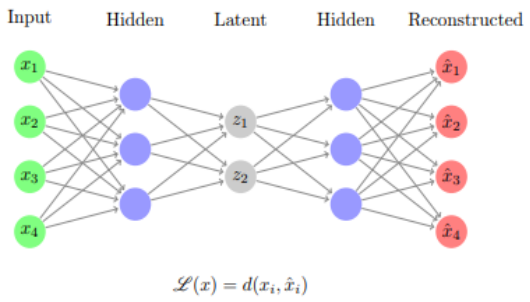

Fig. 1. Autoencoder

reduces the calculation of the main network requirement and the noise of the original inputs.

The learned characteristics are fed to the building and training of models. 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 fig 3. 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.

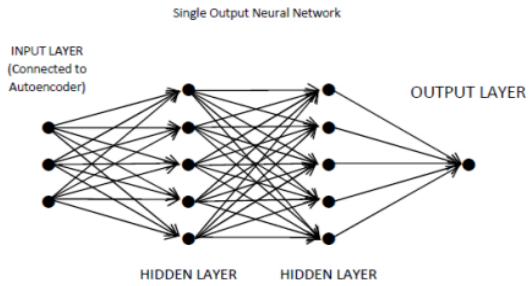

Fig. 2. Neural network

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

V. EXPERIMENTS

This Section shows acquired results of data exploration .Tables IV and V represents missing values in train and test dataset of Life Insurance dataset. Tables VI and VII represents missing values in train and test dataset of Life Insurance dataset.Tables VIII and IX represents missing values in train and test dataset of Life Insurance dataset.As shown every dataset has a severe problem of missing values and required to perform various imputation techniques to make data qualitative. Barcharts Fig 4 and Fig 5 shows the results of Inbalanced target distribution check and As observed Barcharts are not properly distributed which leaded to data imbalance and proper measures as explained in methodology should be taken which will be shown in next paper.

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 IV

MISSING VALUES IN LIFE INSURANCE DATASET TRAIN DATA

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 V

MISSING VALUES IN LIFE INSURANCE DATASET TEST 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 VI

MISSING VALUES IN DRIVER INSURANCE CLAIM DATASET TRAIN 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 VII

MISSING VALUES IN DRIVER INSURANCE CLAIM DATASET TEST DATA

features	Missing Count	missing count percentage
Blind Model	8431	0.064
Blind Make	8431	0.064
Blind Submodel	8431	0.064
Cat1	25981	0.197
Cat2	4874164	3.697
Cat3	3999	0.030
Cat4	5631649	42.7
Cat5	5637321	42.7
Cat6	25981	0.197
Cat7	7167634	54.4
Cat8	3364	0.026
Cat10	3917	0.029
Cat11	31469	0.239
Cat12	28882	0.219

TABLE VIII

MISSING VALUES IN CLAIM AMOUNT PREDICTION DATASET TRAIN DATA

features	Missing Count	missing count percentage
Blind Model	8431	0.064
Blind Make	8431	0.064
Blind Submodel	8431	0.064
Cat1	25981	0.197
Cat2	4874164	3.697
Cat3	3999	0.030
Cat4	5631649	42.7
Cat5	5637321	42.7
Cat6	25981	0.197
Cat7	7167634	54.4
Cat8	3364	0.026
Cat10	3917	0.029
Cat11	31469	0.239
Cat12	28882	0.219

TABLE IX

MISSING VALUES IN CLAIM AMOUNT PREDICTION DATASET TEST DATA

[!ht]

VI. DISCUSSIONS

A. Evaluation

1) Normalized Gini Coefficient: The specified evaluation measure for the Kaggle contest is the standardized Gini coefficient. Statistical response models have traditionally

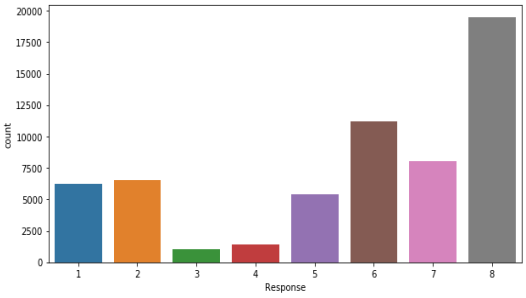

Fig. 3. Target distribution in Life Insurance Dataset

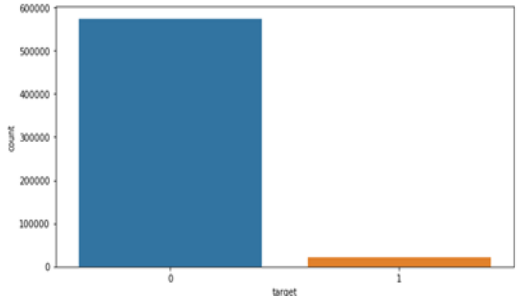

Fig. 4. Target distribution in Driver Insurance Claim Dataset

been assessed on the basis of some form of fitness. Assumptions are made concerning the basic distribution of data and models are evaluated on the basis of how well predicted values fit the observed data values from a sample data set. For the evaluation or production of fitness, various statistical measures (likelihood, R 2), F statistics, Chi Square statistics, classification indices, etc.) are used. In above three datasts, the Kaggle contest required the contestants to use the normalized Gini coefficient as the evaluation measure.

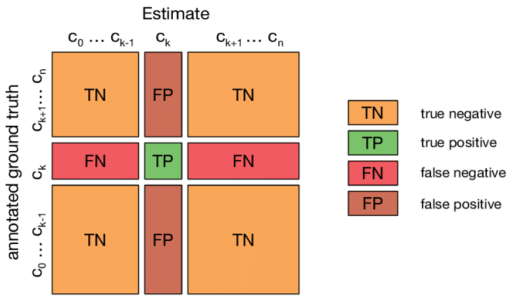

Fig. 5. Confusion matrix

2) 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)

VII. CONCLUSIONS

The primary focus of this paper is to perform task of data set selection and extract some data insights by vari-

ous exploration techniques . Data has been collected form Kaggle webiste under insurance domain.Up on performing various exploration techniques we have observed data has a severe problem of missing values ,outliers and imbalanced target distributions. Various methods has been discussed to overcome those problems .Upon completion of data exploration there are various tasks to be performed to accomplish objective which was discussed in furthur work section will be continued in next paper.

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

Plan Of Action

Furthur Plan Fig 7 shows completed, inprogress and not started tasks of entire objective. As shown dataset collection and exploration has been completed .Data-preprocessing steps are completed includes Missing value imputaion, handling outliers and Feature Engineering, where Modelling step is in progress and Result Evaluation is going to start soon

At Risk	Task Name	Status	Start Date	End Date
	Dataset collection and data Exploration		**02/01/19**	**02/21/19**
	Dataset Set collection	Complete	02/01/19	02/15/19
	Data Exploration	Complete	02/15/19	02/21/19
	Data Pre-processing		**02/21/19**	**03/21/19**
	Missing values Imputation	Complete	02/21/19	02/28/19
	Handiling outliers	Complete	02/28/19	03/07/19
	Feature Engineering	In Progress	03/08/19	03/21/19
	Modelling		**03/23/19**	**04/06/19**
	Feature extraction using Auto encoder	In Progress	03/23/19	03/31/19
	Training standard discriminative neural network with Learned features	In Progress	03/31/19	04/06/19
	Evaluation	**Not Started**	**04/07/19**	**04/20/19**
	Comparing AUC scores	Not Started	04/07/19	04/20/19
	On Sample data	Not Started	04/07/19	04/14/19
	Progressively Increased data	Not Started	04/15/19	04/20/19
	Comparing Results with SAT Methods	Not Started		04/20/19

PLAN OF ACTION

Furthur Plan Fig 7 shows completed, inprogress and not started tasks of entire objective. As shown dataset collection and exploration has been completed .Data-preprocessing steps are completed includes Missing value imputaion, handling outliers and Feature Engineering, where Modelling step is in progress and Result Evaluation is going to start soon

At Risk	Task Name	Status	Start Date	End Date
	Dataset collection and data Exploration		**02/01/19**	**02/21/19**
	Dataset Set collection	Complete	02/01/19	02/15/19
	Data Exploration	Complete	02/15/19	02/21/19
	Data Pre-processing		**02/21/19**	**03/21/19**
	Missing values Imputation	Complete	02/21/19	02/28/19
	Handiling outliers	Complete	02/28/19	03/07/19
	Feature Engineering	In Progress	03/08/19	03/21/19
	Modelling		**03/23/19**	**04/06/19**
	Feature extraction using Auto encoder	In Progress	03/23/19	03/31/19
	Training standard discriminative neural network with Learned features	In Progress	03/31/19	04/06/19
	Evaluation	**Not Started**	**04/07/19**	**04/20/19**
	Comparing AUC scores	Not Started	04/07/19	04/20/19
	On Sample data	Not Started	04/07/19	04/14/19
	Progressively Increased data	Not Started	04/15/19	04/20/19
	Comparing Results with SAT Methods	Not Started		04/20/19