Approaching Machine Learning Using H2O and Inspecting it on Apparel Industry Using D3 Visualization

Chaitra D.B, Mrs.Bindiya M.K PG
SCHALOR, Associate Professor
Dept. of Computer Science and Engineering
SJB Institute of Technology Bangalore-60, India
chaitradbhogesh@gmail.com, bindiyamk2004@yahoo.com

ABSTRACT- Implementing machine learning algorithms was never a easy task, Although there are several approach to do so, but the procedure followed is always exasperating. In order to simplify the task of analyses with same accuracy we give a new approach of H2O.H2O is fast ,scalable, distributed, machine learning and deep learning application .It is a smarter Which implements many generalized linear models like linear regression, logistic regression ,Naive Bayes, K-means clustering, Naive Bayes algorithms .Here in this paper we are trying to approach H2O machine learning with more complex algorithms with an effective solution which can be which can be implemented on various real time problems, one such among them is apparel industry which still follows a traditional inspection and analysis system so using machine learning and D3 visualization we are trying to give better analyses and recommendation to the system.

Keywords -Machine Learning, H2O, Analytics, GLM algorithm, D3 visualization.

I. INTRODUCTION

Over the past few decades Machine Learning has become one of the important part of technology. With the ever increasing amounts of data and which gives rise to a good reason to believe that smart analysis will become even more necessary ingredient for the technological process.

Machine Learning deals with automating the automation that is making the machine intelligent [1].Using the past example data or using the past experience to solve a given problem many successful applications have been developed one such application is H2O [2].

A. H2O Machine Learning

H2O is one of the machine learning application which makes it easy for anyone to apply predictive analysis and math to solve today's any of the most challenging business problems. Many features of machine learning have been intelligently combined this platform which is nor currently present in any other machine learning platforms.

In H2O it is easy to combine the power of different highly advanced algorithms and the truly scalable in memory processing capacity for big data to increase the benefit for the business [3]. With H2O it is possible to make better predictions, ready to use algorithms can be harnessed and also the super power which is needed to analyze bigger data sets, more variables and models can be easily obtained. With H2O we can work with any of existing languages and also tools and can even expand the platform seamlessly to our Hadoop environments.

H2O uses familiar programming environments like Python,R, JSON,Scala and JAVA through H2O's powerful API's.Using H2O we can explore the model of any big data from within R Studio,Microsoft Excel and Tableau and many more.Through H2O it is easy to connect data from NoSQL, S3, HDFS and SQL and other data sources and H2O can be installed on any platform and deployed also.

H2O supports Nano Fast scoring engine and also we can train any number of data models and can iterate over those models to accurate models in real time using H2O's in memory distributed parallel processing develop.

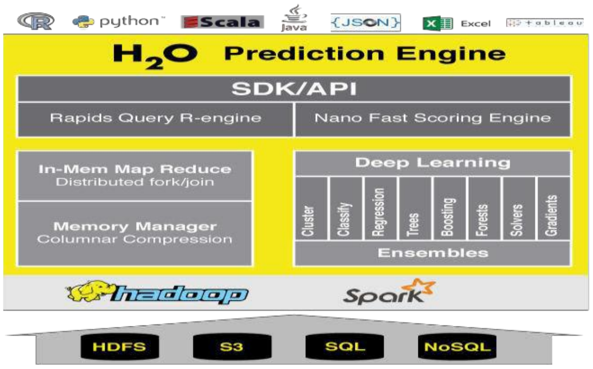


Figure 1: H2O Machine Learning Architecture

Once the analysis of the data has been done using the H2O machine learning we do the visualization of the data using the D3 [4].

B. D3 visualization

D3 gives us the liberty to bind any arbitrary data to a DOM that is Document Object Model and then we apply the data driven transformations to those documents [5].D3 a data driven document is just not a monolithic framework that only provide every conceivable feature, but also solves the crux of the given problem and does the efficient manipulation of the documents based on the data contained in that particular document. This quality of D3 avoids the proprietary representation of data by enabling the extraordinary flexibility and also exposes the web standards like SVG,HTML and CSS .D3 is a functional style representation which can be modified according to the needs.D3 with the little overhead is extremely fast and it supports any range of large datasets and most important thing of D3 is that it gives the dynamic interaction to the graphs and animations because of which richness is added to the D3 visualization .

II..RELATED WORK

Work on "Social Contextual Recommendation" that there is an epidemic growth in the data by social media which is cannot be handled by the traditional database systems. The existing tools consider the social structure of network but not the social context in full so recommend the things hence he proposes a model which does the recommendation based on the sociology study of the person on his different social media activities and his individual likes and dislikes his preferences to do the recommendations. On the basis of his interpersonal influences also recommendation is done. The empirical results are based on the datasets collected from Facebook, Twitter.

Paper "Layer-Cantered Approach for Multigraphs Visualization" he demonstrates that recent advances in the social media, network sciences involves the creation of models and analysis of generated data. . In this paper he demonstrates the new model for multiple edge graphs and to manage the different types of relationship between existing multigraphs. In this approach the pair of edges or nodes are specifically studied, and how the different nodes are interacted together based on the edges they share. In this paper in two level of global and local multigraph features are considered. The global approach deals with the gaining the knowledge which are related to the different characteristics of layer they combine with, where in the local gives the analysis of separate layers and explain each characteristics separately. This proposal is tested with real world data and it has helped to bring about useful patterns.

III. MOTIVATING EXAMPLE

Big data analytics is picking up a large data sets of high volume and high variety of data because of which it is called big data and processing it for business intelligence. There are many customer preferences, hidden patterns, unknown correlations and also many other business information which are useful for analysis.

In many social media websites they use re commander systems to recommend the new movies, songs, books and many other things based on the correlation of the members in the social media .If two persons like something in common along with their other personal unique likes system tries to recommend those unique likes also to each individual who share few things in common thinking that they may like these recommendations also. This is how the market promotion is done for items these days .Association rules also plays important roles in this recommendation systems like if someone buys phone he may be recommended with memory cards likewise.

Using the quite a same is expanded to apparel industry ,there are various band apparels now a days all the apparels manufacturer does not need to own his own industry he may give his orders to other local industries who take the orders from any customers and develop their needs and supply to them.

The interesting part of this process is inspecting the developed products from the suppliers by the inspectors .There may be any kind of mistakes from the suppliers end like the material of the cloth may be different or accessories used, quality of the product, size of the product likewise there can be any number of defects or even it may be perfect .The quality manager or the inspector may find and need to generate report. Thereby huge number of reports are being generated across the globe.

Thus we have taken all these reports and analyzed for the quality of each supplier to each customer for all possible years and found who can be the most reliable supplier ,what are his defects rates, is that increased or decreased compared to older years and predicting about its future years rate. By taking his overall performance we can recommend to the customer whether it's worth giving order to that particular supplier or not. All this analyzed data is represented in eye catchy D3 visualization techniques so that it's easy of common people to understandable and consider the recommendation.

IV. THEAPPROACH

We approached the problem of apparel industry through the concepts of machine learning and used other technologies like NO-SQL, MongoDB, JSON, SQLite and also visualization technique called D3.We have approached machine learning here using H2O platform. First considering the reports generated by the inspectors ,collected all the reports in which the format they are generated that is either in the form of CSV ,Microsoft excel or in any other format. The data inside these documents is the input data and this data is used for the training the systems.

The reports irrespective of the form they are generated they are they are converted into standard JSON which is first manually created and checked for the universality for all other reports and once confirmed it is automated .Any new report comes it automatically converted into the JSON structure defined .Then the dump of SQL tables are generated using which JSON files are created ,for every report there is single file generated for the purpose of extracting the required data and then moving the files to MongoDB which is a NO-SQL database for the further process of analysis.

Once the data is available on the MongoDB it passed on to the H2O machine learning platform using R. Then the data is analyzed using the GLM (Generalized Linear Model) which is the algorithm for both classification and regression using which when a new supplier for the customer comes in it decides whether that supplier is the high ,low or moderate customers for that customer that is classification is done .Then after analyzing all the data the visualization is done using the D3 graphs these are the highly interactive graphs wherein a person can easily understand the analysis done without missing the single part.

A.The GLM algorithm

GLM is one of the flexible model in statistics, it is the generalization of the ordinary linear regression model. GLM is for both classification and regression [6]. GLM estimates the regression for the exponential distribution values. GLM along with the Gaussian distribution it also supports Poisson, Gamma, Binomial and also Tweedie distributions

Generalized GLM is given by relating the linear model via a link function and by taking the range of the variance for each measurement to be a function of its predicted value. GLM allows response variables that have error distribution other than normal distribution [7] [8] [9].

For the estimation of the model ,an iteratively reweighed least square method is considered for maximum likelihood .GLM were form as a way of unifying other statistical models like logistic regression, linear and poisson regression. Least square fits and Bayesian approaches to variance stabilized responses also have been developed.

he GLM suite includes

- Gaussian regression
- Gamma regression
- Poisson regression
- Tweedie regression and
- Binomial regression

C. Usage

h2o.glm(x, y, data, key = "", offset = NULL, family, link, tweedie.p = ifelse(family="tweedie", 1.5),strong_rules = TRUE, alpha = 0.5, prior = NULL, standardize = TRUE,beta_constraints = NULL, nfolds = 0 , use_all_factor_levels = FALSE, lambda_search = FALSE, , disable_line_search = FALSE,nlambda = -1, max_predictors = -1, return_all_lambda = FALSE, intercept = TRUE,lambda.min.ratio = -1, non_negative = FALSE,variable_importances = FALSE, iter.max = 100, higher_accuracy = FALSE)

D.Variables of GLM Model

Response: It is the model dependent variable and is denoted as Y .While choosing the appropriate model for distribution for estimating specific features of a dependent variable should be taken into consideration

Gaussian: Y must be real valued and contiguous variable

Gamma: Y variable must be valued strictly greater than 0 and also it must be discrete

Binomial: Y variable are valued only at 0 or 1 and also discrete

Tweedie: Y variable are combination of Poisson-Gamma compound distribution

Poisson: It is used as a model count data. These Y variables are strictly greater than 0.

Ignored Columns: From the current data set a list of columns field will be auto-populated.H2O gives the liberty of selecting the set of columns which can be omitted from the processing model.H2O by itself omits the column with the constant values and also dependent variable in Y. Since variances are constant columns are omitted.

Standardize: Variables are transformed into standardized variables, each with unit variance and mean value of 0.Co-efficients and Variables are expressed with respect to their standard units and relative 0 position.

Maximum iteration: When the data set is given to the system, these datasets are iteratively processed to increase the accuracy. As the number of iteration increases accuracy also increases. If maximum iteration is set to 100 then algorithm repeats the gradient descent 100

times .These iterations are performed for training examples. N folds, this is the number which specifies cross-validation models to generate parallel for training a model on the whole data set. If value of Nfolds is set to 10, additional models are generated with 1/10 of data for each data.

Family and Link: In H2O GLM function each family is connected with a default link function, each function defines the specialized transformation on the set of variables that is X variables chosen to predict Y.

Gaussian: It is the identity here the Y are contiguous or discrete predicted values and quantitative that can be interpreted as approximately continuous.

Gamma: It is the inverse, and even distributed as Poisson, here the variance is greater than mean.

Binomial: It is the logit value. These binomial dependent variables are taken as categorical Y with two possible outcomes 0 and 1.

E.Validate GLM:

Cross Validation: GLM analyses in H2O are presented with cross-validation models, here the coefficients presented in the model are not dependent on the cross-validation models, and the coefficients are generated with least-squares on the whole dataset. Cross validation values are generated by taking 90% of sub sample data and remaining 10% with training and test data [10].

Cost of Computation: H2O gives a distributed parallel computing hence large amount of data can be processed, here the large sets are divided into smaller sets and then processed [11].

In GLM, data are divided into rows and not columns, this is because the Y value are dependent on information of predicator variable vectors.

If O is taken as complexity function and P as number of columns or predictors and N as the number of rows or observation then

Equation 1: $Runtime \propto p^3 + \frac{(N * p2)}{CPU_s}$

By above formula it is evident that distribution reduces the time it takes to process the algorithm as it decreases N.

LM Algorithm

Let $y_1, \cdots\cdots y_n$ be n observations of the independent, random response variable y_i [13] [14].

Assume that the observations are distributed according to a function from the exponential family and have a probability density function of the form:

Equation2: $f(y_i) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i;\phi)\right]$

Where θ and ϕ are location and scale parameters. And $a_i(\phi), b(\theta_i), c(y_i;\phi)$ are known functions. a_i Is of the form: $a_i = \frac{\phi}{p_i}; p_i$ is a known prior weight. When Y has a pdf from the exponential family:

$$\vee(\theta_i)a_i(\phi)$$

$$E(Y_i) = \mu_i = b \; var(Y_i) = \sigma_i^2 = b$$

Let $g(\mu_i) = \eta_i$ be a monotonic, differentiable transformation of the expected value of y_i. The function η_i is the link function and follows a linear model.

$$g(\mu_i) = \eta_i = x_i' \beta$$

When inverted: $\mu = g^{-1}(x_i' \beta)$

G .Other technologies used

Mongo DB: Mongo DB is one of the widely used NoSQL database .It is cross-platform and document oriented document database. MongoDB is widely used since it entangles the traditional table based data that is relational database in the favor of JSON like documents.

MongoDB has features like AD hoc queries, Load balancing, Replication, File storage, importantly aggregation and many other.

It also gives the high performance, availability and automatic sharding. The records are stored in the form of key and value pairs. The value of the filed maybe any other documents or arrays or even the arrays of documents also. Since all the data are in form of objects it is easy to access and analyze .It supports horizontal scalability and also multiple storage engines.

JSON: It is the abbreviated form of JavaScript Object Notation, as the name itself specifies the data are stored in the form of key and value pairs and values can be again a document or arrays. It is very easy for humans to read and write this kind of format .It is in the text format and fully independent of the language.

JSON is built in two ways, one is a collection of key and value pairs like record, object, associative array and other is ordered list of values like vector, array, list etc. JSON is a unordered set of data. A data is embraced in open and close braces.

V. ALGORITHMS

A. Algorithm for Creating JSON

Step 1: Gathering the required dump from client, dump contains the data of inspected reports which can be in any format.

Step2: Converting the dump into SQL tables using SQLite

Step3: Processing the tables to generate JSON, for each data separate JSON

Step4: Cleaning the JSON and pre-processing for convenient or required data.

Step5: Loading the obtained resultant JSON to MongoDB

B. Algorithm for processing the data inside MongoDB

Step1: Create an API for data extraction from JSON in MongoDB

Step 2: Process the extracted data into machine readable JSON format output

Step3: Pass the JSON output to the GLM algorithm for analysis

Step 4: compare the results of all the datasets previous and current if produced result is acceptable then accept else repeat step 2 and 3

Step 5: once output is acceptable pass it to D3 charts for visualization

C. Algorithm for Visualization

Step 1: Pass the output result of GLM algorithm to D3 charts for generating graphs

Step 2: Collect feedback from D3 charts to make it interactive, if interaction works correctly it is accepted or else change the blocks used for interaction

VI. EXPECTED RESULTS

In this section we demonstrate the expected output of our approach alongside comparing the existing model results. Specifically, we demonstrate on working of algorithm: 1) creating the JSON 2) loading data to the database 3) analyzing the data using H2O and 4) visualization using D3 chats.

A. Creating JSON:

In the creation of the JSON here we create one single standard JSON .This JSON structure is common for all reports data any data coming in will be converted to this standard JSON format which is in the form of key and value pairs which is easy to understand and access for further analyses .Whereas the existing model contains the report in the way they are generated like Microsoft Excel ,CSV of any other format which is difficult to access and analyze and get the relevant result .Different reports are in different manner and it is tedious to match the compatibility of the reports and because of which the accuracy of the result goes down. Hence the conversion of all the reports to single standard format helps to obtain the fast processing and also accurate result.

B. Moving JSON to MongoDB:

After creating the standard JSON structure it is moved to the No-SQL database MongoDB. MongoDB is a schema less database because of which we can take a liberty of changing the report formats .The data coming in may contain the missing data ,redundant data or even the data which is not useful for the process ,taking all this parameters MongoDB is the best database as it by default gives all these features and aggregation is done which is required for the analyses .MongoDB gives the best aggregation queries to be used and in our result it gives the various aggregation result like according to the aggregating the suppliers and customers which supplier quantity of the items supplied all these aggregations are easily done again which is very much required for analyses process. Whereas which is not possible in existing system which uses the relational database Which is schema dependent database and using which dynamic schema cannot be generated.

C.Analysis:

After passing all the data to the MongoDB, considering only relevant data analyses is carried out. Data contains the information about the client ,supplier, location , quantity of items built and supplied to the customer and what are the major defects ,minor ones and which needs to be prioritized first and all the analytics is done here all these analyses is done using the GLM algorithm and this algorithm gives the accurate result. By using the results of the analyses one can find the moderate, average and low suppliers and based on this analyses companies can decide to which supplier it can consider next. Where there is no analyses done in the present system.

D. D3 Visualization:

Once the analyses result is got it can be visualized using D3 chats which are very interactive and easy to understand the result. These interactive graphs gives the clear picture of overall analyses done about client, supplier, quality ,quantity, their consistency range over the years .Here we are generating the stacked bar chart ,Sequential Sunburst, Pie charts and line charts.

In the stacked bar graph we compare the supplier and the quantity of items he has delivered to his customers over a period of time. This gives the consistency analyses of the supplier over range of years. This graph is made interactive by taking either staked or grouped result.

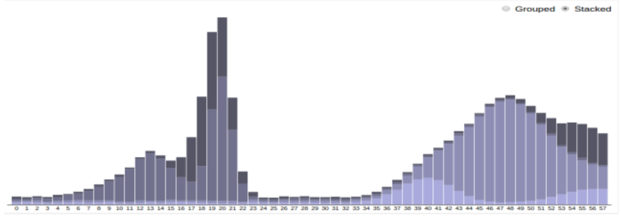


Figure 2: Stacked bar graph

Using the pie charts we represented who is the low ,medium and high suppliers over a period of time .Here based on the percentage of the mid, high and low pie chart is distributed .It is made highly interactive by clicking on one area of pie charts it gives pop of another bar chart with relative values.

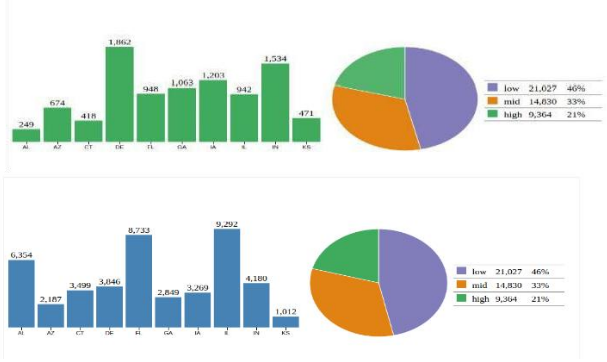


Figure3 a) Pie chart for low range b) Pie chart for mid-range

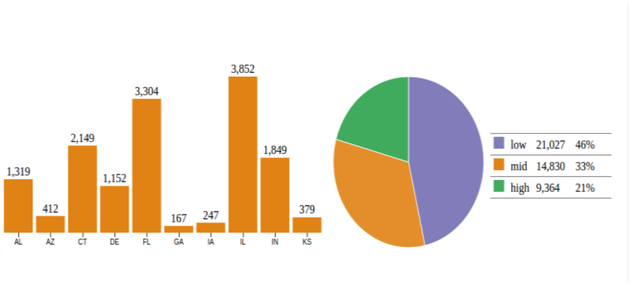


Figure 4: Pie chart for high range

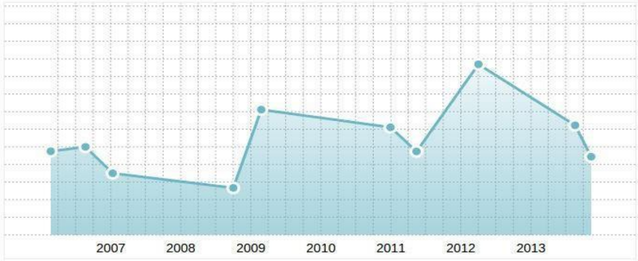


Figure5: Area graph over range of years

Area graph is used to give the performance of the particular supplier over a period of time or years.

Overall analyses can be even represented in a D3 chart called sequential sunburst. This graph gives the information about the client, supplier, location of the industry, name of the industry and all other values. By clicking on one part of the chart it redraws and takes the other values according to the requirement.

Figure6: Sequential sunburst for overall model

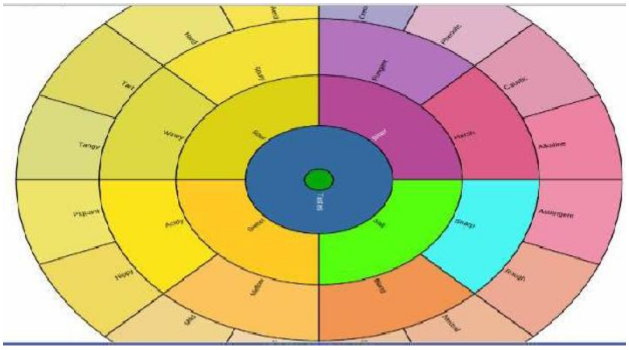


Figure7: Sequential sunburst when one level reduced

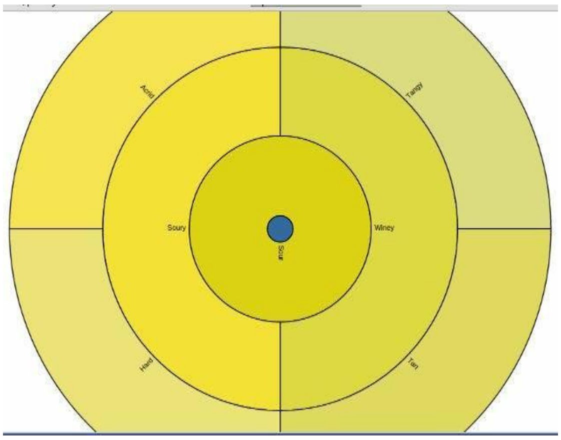
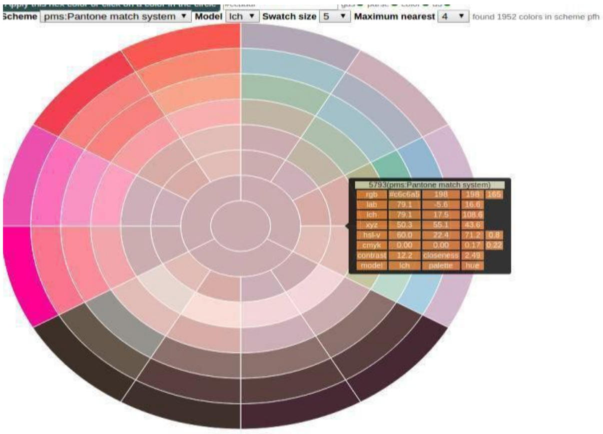


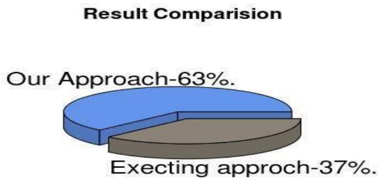
Figure8: Sequential sunburst when two level reduced


Scheme pms:Pantone match system ▾ Model lch ▾ Swatch size 5 ▾ Maximum nearest 4 ▾ found 1952 colors in scheme pfh

5793(pms:Pantone match system)

rgb	#c6c6a5	198	198	165
lab	79.1	-5.6	16.6	
lch	79.1	17.5	108.6	
xyz	50.3	55.1	43.6	
hsl-v	60.0	22.4	71.2	0.8
cmyk	0.00	0.00	0.17	0.22
contrast	12.2	closeness	2.49	
model	lch	palette	hue	

Figure9: Sunburst with popup table information



CONCLUSION

In this paper we have tried to analyze the data of apparel industry and give them a better approach to analytics using H2O machine learning application. We have tried to convert all the reports to one single standard structure of JSON and successfully stored to MongoDB and from where the analysis is done for the better supplier and customer and quality of the product which is recommended for the people and also future predictions and customer performance is found. At the last all the analysis is represented in the form of charts for the better understanding. Bylooking into the graph one can predict how is the moderate, low or high supplier and also customer, quality of the product, quantity of the product supplied and what can be his future quantity and quality and where is the location of the industry, who are all the customer for which customer and other possible analytics of all the apparel industry present across the globe. The results obtained were better promising then the existing model.

REFERENCES

[1] http://www.sas.com/en_us/insights/analytics/machine-learning.html

[2] https://www.r-project.org/about.html

[3] http://h2o.ai/product/algorithms/

[4] http://searchbusinessanalytics.techtarget.com/definition/data-visualization

[5] https://d3js.org/

[6] http://h2orelease.s3.amazonaws.com/h2o/master/1713/docs-website/datascience/glm.html

[7] Breslow, N E. "Generalized Linear Models: Checking Assumptions and Strengthening Conclusions," Statistica Applicata 8 (1996): 23-41.

[8] Frome, E L, "The Analysis of Rates Using Poisson Regression Models," Biometrics (1983): 665-674. http://www.csm.ornl.gov/~frome/BE/FP/FromeBiometrics83.pdf

[9] Goldberger, Arthur S, "Best Linear Unbiased Prediction in the Generalized Linear Regression Model," Journal of the American Statistical Association 57.298 (1962): 369-375.

[10] Guisan, Antoine, Thomas C Edwards Jr, and Trevor Hastie, "Generalized Linear and Generalized Additive Models in Studies of Species Distributions: Setting the Scene," Ecological modeling 157.2 (2002): 89-100. http://www.stanford.edu/~hastie/Papers/GuisanEtAl_EcolModel-2003.pdf

[11] Nelder, John A, and Robert WM Wedderburn, "Generalized Linear Models," Journal of the Royal Statistical Society. Series a (General) (1972): 370-384. http://biecek.pl/MIMUW/uploads/Nelder_GLM.pd f

[12] Niu, Feng, et al. "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," Advances in Neural Information Processing Systems 24 (2011):693-701.http://www.eecs.berkeley.edu/~brecht/papers/hogwildTR.pdf