ALLSTATE CLAIMS SEVERITY

USING R

**INDEX:**

|  |  |
| --- | --- |
| **PROBLEM DISCUSSION** | Pg.1 |
| **SIGNIFICANCE** | Pg.2 |
| **EXPLORATORY ANALYSIS** | Pg.2 |
| **DISCUSSION ON MODELS** | Pg.4 |
| **LITERATURE** | Pg.5 |
| **FORMULATION** | Pg.5 |
| **MODEL PERFORMANCE** | Pg.6 |
| **LEARNING AND LIMITATIONS** | Pg.6 |

**1.Problem Discussion**: Insurance companies are continously seeking ways to improve the customer experience right from pushing paper to an insurance agent till the claims are received . Allstate ,a personal insurer in United States, is continually seeking fresh ideas to improve their claims service for the over 16 million households they protect. They are have plans to develop automated methods to predict the cost(loss incurred).

***Objective*:**

1. Understanding the training data provided by AllState
2. After features are understood, developing a prediction model to predict the “loss” variable
3. Examine the errors and tune the model.

**2.Significance**:  Automating the cost(loss incurred) prediction might have many applications, including

* Identification of factors that have major impact on **loss**.**[Loss** of Use Insurance - policy providing protection against loss of use due to damage or destruction of property**]**
* Ensure a worry-free customer experience

**3.Exploratory Analysis / Data Cleaning**.

There are 188319 rows and 132 columns in the training set. And there are 125547 rows and 131 columns in the test set.There are **no missing values** in both the training and the test sets. Each Claims are uniquely identified by an ‘id’ and they are represented by 116 categorical variables and 14 continuous variables.

We box-plotted the 14 continuous variables(cont1,cont2…) and it happens to be that variables cont7 and cont9 have outliers. Since only 2 variables have outliers we decided to ignore them in the assumption that it will have only a negligible impact on the model.And there are no missing data.

There are 125547 rows in the test set which is almost 75% of the number of rows in the training set .Predicting the 125547 values in ‘loss’ variable requires a very accurate prediction model

***Data:***

* **Id:**This variable is used to uniquely identify each claim.
* **Cat1-Cat116:**117 Categorical variables per claim
* **Cont1-Cont14:**15 Continuous variables per claim

***Missing Data:***

There are no missing values in the data provided by AllState.

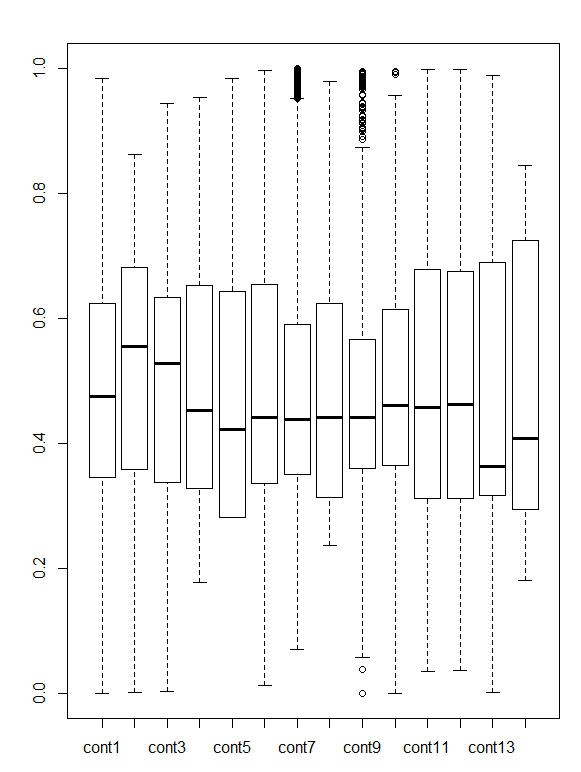




he default method for **is**.**na** applied to an atomic vector returns a logical vector of the same length as its argument x , containing TRUE for those elements marked **NA**or, for numeric or complex vectors, NaN , and FALSE otherwise. (A complex value **is**regarded as **NA** if either its real or imaginary part **is NA** or NaN .). Summing those will give the total number of NAs.

***Outliers:***

There are outliers in cont7 and cont9 variables .The box plot below proves that.



**4.Discussion on Models**:

First analysis we need before deciding on the model was to analyse the nature of the data. The data given to us has 116 categorical(Nominal) and 14 Continuous variables(Numerical).Therefore this kind of data falls under Mixed Data.

After certain Research we indentified that tree-based models work well with mixed data. The first tree-based model that came to our mind was decision tree .Before applying tree-based models we thought of applying dimension reduction techniques.

We pondered over using the following dimension reduction methods

**MCA(Multiple Correspondence Analysis):** In statistics, **multiple correspondence analysis** (**MCA**) is a data analysis technique for nominal categorical data, used to detect and represent underlying structures in a data set. It does this by representing data as points in a low-dimensional Euclidean space. The procedure thus appears to be the counterpart of principal component analysis for categorical data.[[1]](https://en.wikipedia.org/wiki/Multiple_correspondence_analysis#cite_note-1)[[2]](https://en.wikipedia.org/wiki/Multiple_correspondence_analysis#cite_note-2) MCA can be viewed as an extension of simple correspondence analysis (CA) in that it is applicable to a large set of categorical variables.

**FAMD(Factor Analysis on Mixed Data):** In statistics, factor analysis of mixed data (FAMD), or factorial analysis of mixed data, is the factorial method devoted to data tables in which a group of individuals is described both by quantitative and qualitative variables. It belongs to the exploratory methods developed by the French school called *Analyse des données* founded by Jean-Paul Benzécri.The term *mixed* refers to the simultaneous presence, as active elements, of quantitative and qualitative variables. Roughly, we can say that FAMD works as a principal components analysis (PCA) for quantitative variables and as a multiple correspondence analysis (MCA) for qualitative variables.

To overcome the above problem we planned to apply random forest model to data as it has inbuilt dimension reduction capability. We used Random forest H2o as it is faster compared to Random Forest under “RandomForest” Package.

In order to improve performance of weak tree based models ensemble approach has been used. Ensemble modelling can be approached in three different ways:

1. Bagging –RandomForest
2. Boosting – GBM ,XgBoost
3. Stacking

The techniques of bagging and boosting have been used in this scenario inorder to improve model performance.

**Bagging**: Trains each model in the ensemble using a randomly drawn subset of the training set reducing the variation of the prediction model. Data is sampled at random along with random predictors and average outcome of all the models is considered like in Random Forest.(Parallel)

**Boosting:** It involves building a model and re classifying its erroneous classifications by some cost function hence boosting the model performance.(Cascade)H2o.gbm has been used to implement boosting.

**5.Literature**:

* **Multiple Correspondence Analysis:**
* An article on Statistical Tools for Higher-throughput Data Analysis(STHDA) on Multiple Correspondence analysis talks in detail about performing and interpreting MCA.The article shows how to use FactoMineR package for **computing Multiple correspondence analysis** (MCA).It is a step-by-step guide for **interpreting** and **visualizing** the output of MCA.

**Ref:**<http://www.sthda.com/english/wiki/correspondence-analysis-in-r-the-ultimate-guide-for-the-analysis-the-visualization-and-the-interpretation-r-software-and-data-mining>

* Lecturer in the Department of Statistics at the University of California Berkeley,Gaston Sanchez has website where he frequently posts blogs on data analysis and statistics.One of his blogs that was related to MCA was “**5 functions to do Multiple Correspondence Analysis in R”.** It clearly explained 5 different packages and functions in R to perform Multiple Correspondence Analysis

**Ref:** <http://gastonsanchez.com/how-to/2012/10/13/MCA-in-R/>

* **Factor Analysis on Mixed Data:**
* Two articles were primarily useful in understanding and implementing factor analysis on mixed data.They not only explained the physical meaning of **FAMD** but also on the implemention of famd algorithm in R.

**Ref:**<https://www.rdocumentation.org/packages/FactoMineR/versions/1.34/topics/FAMD>

**Ref:** <https://rdrr.io/cran/FactoMineR/man/FAMD.html>

* **Dealing with Categorical Variables:**
* There was an article in analyticsvidhya.com on the topic “Simple Methods to deal with Categorical Variables in Predictive Modeling”.It explained 3 methods to deal with categorical variables-Convert To Number,Combine Levels and Dummy Coding.Even we did not convert the categorical variables the article provided some useful information on how to deal with categorical variables.

**Ref:**[https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling](https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/)**/**

* There was article in Kdnuggets.com on **One-Hot Encoding.** It explained how to perform one-hot encoding technique on categorical data. The article was so clearly structured that it explained in simple steps to perform one-hot encoding.

**Ref:**[http://www.kdnuggets.com/2015/12/beyond-one-hot-exploration- categorical-variables.html](http://www.kdnuggets.com/2015/12/beyond-one-hot-exploration-%20%20%20%20categorical-variables.html)

* **H2O Documentation**

The h2o documentation comes with detailed information about

1. Package download
2. Installation
3. Initialization
4. Importing data
5. Scaling
6. Model specific methods and parameters
7. Detailed description of parameters
8. Grid search for parameter tuning.
9. Cross validation and nfolds

Basic examples provided with further research about concept of a particular algorithm helped in selecting and applying them to our model.

**Ref**: <https://h2o-release.s3.amazonaws.com/h2o/rel-turing/1/docs-website/h2o-docs>

**Ref:**<https://www.rdocumentation.org/packages/h2o/versions/3.10.0.8/topics/h2o.gbm>

* **Tree-Based Modelling Techniques**

Detailed explanation of concepts related to tree.

1. Inital introduction to ensemble modelling.
2. Grouping weak learners to form a stronger model
3. Explanation regarding greedy approach in tree
4. Brieman rules for approximating mtries
5. Relation between mtries and error rate
6. Greater co relation among the predictors leads to greater error rate.
7. Categorical decision tree and Regression decision tree.
8. Pruning in tree based algorithms.
9. Removing layers based on negative gain
10. GBM pruning needs to be manually determined
11. Xgboost takes tree to its maximum and traces backwards facilitates pruning

**Ref:**<https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python>

**Xgboost documentation:**

**Ref:**http://xgboost.readthedocs.io/en/latest/R-package/xgboostPresentation.html

REF:https://github.com/dmlc/xgboost/blob/master/R-package/vignettes/discoverYourData.Rmd

1. Package download
2. Installation
3. Initialization
4. Importing data
5. Data conversion into matrix format required for xgboost
6. Model Analysis and tuning.

**6.Formulation / Libraries**:

* **Model:h20.gbm**

The model is formulated by understanding the concept and functioning of the Gradient Boosted Machines. The variable inputs considered while modelling are:

* Distribution: The distribution has been left to default for the model to decide.
* Ntrees: The number of trees to grown to reduce our error to an acceptable rate
* Max\_depth/Shrinkage: The maximum depth the tree should be allowed to grow
* Min\_rows: The number of rows for a node to be terminal node. Proper selection can reduce the depth.
* Learn\_rate: Learning rate for convergence it is inversely proportional to number of trees.Higher rate makes the model a poor predictor
* Validation frame:Test frame for validation else training frame is used.
* Nfolds: Number of folds the data is to be divided into. Helps in Cross validation.
* **Model:h2o.randomForest**

Random forest in h2o provides more parameters to tune than others and also doesn’t have any limit to number of levels of variables in categorical columns.

* x vector providing input frames
* y provides output column
* training\_frame:The data set containing x and y
* sample\_rate: No. of rows sampled per tree
* max\_depth:Maximum depth tree is allowed to grow
* seed : To maintain same random data
* nfolds: Number of folds of cross validation is to be done
* score\_tree\_interval: Record the performace of tree for every interval
* mtries : Number of variables selected for sampling
* **Model:h2o.DeepLearning**

H2O Deep Learning has many parameters, it was designed to be just as easy to use as the other supervised training methods in H2O. Early stopping, automatic data standardization and handling of categorical variables and missing values and adaptive learning rates (per weight) reduce the amount of parameters the user has to specify. Often, it's just the number and sizes of hidden layers, the number of epochs and the activation function and maybe some regularization techniques

Tuning of number of hidden layers and epoch makes a noticable difference to mean absolute error.

* **Model:h2o.gird**

Grid search is a powerful model provided by h2o escpecially for tuning of hyper parameters. It creates several models at once and helps in comparision of several models.

* Modelname: like “gbm”,”randomForest”
* Hyper\_parameters: List of hyper parameters
* Other: Other constant arguments
* **FactoMineR:**

It is a package in R to exploratory data analysis methods to summarize, visualize and describe datasets. The main principal component methods are available, those with the largest potential in terms of applications: principal component analysis (PCA) when variable.  **multiple correspondence analysis** (**MCA**) is a data analysis technique for nominal categorical data, used to detect and represent underlying structures in a data set. It does this by representing data as points in a low-dimensional Euclidean space. The procedure thus appears to be the counterpart of principal component analysis for categorical data. MCA can be viewed as an extension of simple correspondence analysis (CA) in that it is applicable to a large set of categorical variables.It also has a function to perform factor analysis on mixed data. **factor analysis of mixed data** (**FAMD**), or **factorial analysis of mixed data**, is the factorial method devoted to data tables in which a group of individuals is described both by quantitative and qualitative variables

**6.Model Performance**:

The evaluation for the project is done by calculation Mean Absolute Error of the loss calculated.

***H2o.gbm***:

The intial model evalutaions started with 100 trees and only 80% of the data. Mae:1046.364(train\_data)

Mae:1212.12(Valaidation\_data)

The scores obtained from the tree showed no decrease in error from 90-100th tree hence the number of trees were increases to 350

Mae:1011.64 (train\_data)

Mae:1194.632(validation\_data)

Proper understanding of cross validation and nfolds meant split of data is not required.The score of trees too showed gradual decrease hence the next training was performed with 100% data and 400 tree and variety of hyper tuned parameters.

The best out come also showed that depth of trees need to be low.

Mae:1017.672(train\_data)

Mae:1196.735(validation\_frame)

***H2O RandomForest*:**

Computaitons of random forest were relatively faster for same number of tree, but despite higher number of tree and various parameters the error was never below

mae:1200(train\_data)

mae:1246(validation\_frame)

***H2O Deep Learning*:**

With 2 hidden layers of 500 nodes each and an epoch of 800:

MAE:1202(On Kaggle)

**7.Limitations**:

Model doesn’t attempt to reduce the dimensions of the given data or find patterns in the categorical data to try feature engineering. Although scaling of data has been attempted it isn’t properly compared with the error of unscaled data to gauge the performance. Computational time and memory comes at a high cost which forms a dis advantage to properly tune parameters. The outcome doesn’t vary much with scaling of data.

There is another limitation with using RandomForest Package as it does not accept categorical variables with more than 32 factors.

**8.Learning**:

More trees better prediction,More epochs better learning similarly more projects more learning.

Tree which were considered beginner tools have been cast upon a newlight. Their usefulness with mixed data prediction, increasing efficiency Boosting and Bagging. Tuning parameter with grid search which was end objective of the previous project which is implemented and its usefulness and memory limitaions in case of excessive parameters is understood. Better understanding of the package h2o and its features.