**PREDICTIVE MODELING PROJECT**

(OPIM 5604)

**Project Report**

on

**Classification of Risk in the Insurance Industry**

**Team 1**

Ankit Agrawal

Jessica McGrinder

Praveen Gupta Sanka

Seshi Saraswathi Harianathan

**TABLE OF CONTENTS**

**CHAPTER** **PAGE**

Business Case 3

Description 3

Sample - Generate a representative sample of data 3

Explore - Visualization of the data 4

Modify - Data Modification 6

Model - Machine learning models 9

Assess - Business Application 19

Conclusion 19

**Business Case**

Develop a predictive model that accurately classifies risk using the data that is available from an insurance company. Response has to be classified (or) rated from ‘1 – 8’ based on the input features.

Data set is obtained from the following source (*kaggle*):

[*https://www.kaggle.com/c/prudential-life-insurance-assessment/data*](https://www.kaggle.com/c/prudential-life-insurance-assessment/data)

**Description**

In a one-click shopping world with on-demand everything, the life insurance application process is antiquated. Customers provide extensive information to identify risk classification and eligibility, including scheduling medical exams, a process that takes an average of 30 days.

The aim of this project is to develop predictive power of the data points in the existing assessment, enabling the firms to significantly streamline the process.

A common data mining approach – SEMMA has been followed to classify the risk for this problem. Sample, Explore, Modify, Model and Assess are the five steps involved in the SEMMA approach.

**Sample - Generate a representative sample of data**

There are 59,382 records available in the data set obtained from the Kaggle. This dataset is split in to three parts – train, cross validation and test. A train dataset is a sample used for learning the classification based on the predictor parameters. A cross validation dataset is used to reduce the overfitting in the model. A test dataset is a completely independent dataset on either of the above and on which the target variable has to be predicted.

Stratified splitting of data is done based on ‘Response’ column to generate the aforementioned 3 datasets. Splitting of data is done in the ratio – 50, 30, 20 for obtaining train, cross validation and test datasets respectively.

Following are the features and their description which are available in the data set.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Id | A unique identifier associated with an application. |
| Product\_Info\_1-7 | A set of normalized variables relating to the product applied for |
| Ins\_Age | Normalized age of applicant |
| Ht | Normalized height of applicant |
| Wt | Normalized weight of applicant |
| BMI | Normalized BMI of applicant |
| Employment\_Info\_1-6 | A set of normalized variables relating to the employment history of the applicant. |
| InsuredInfo\_1-6 | A set of normalized variables providing information about the applicant. |
| Insurance\_History\_1-9 | A set of normalized variables relating to the insurance history of the applicant. |
| Family\_Hist\_1-5 | A set of normalized variables relating to the family history of the applicant. |
| Medical\_History\_1-41 | A set of normalized variables relating to the medical history of the applicant. |
| Medical\_Keyword\_1-48 | A set of dummy variables relating to the presence of/absence of a medical keyword being associated with the application. |
| Response | This is the target variable, an ordinal variable relating to the final decision associated with an application |

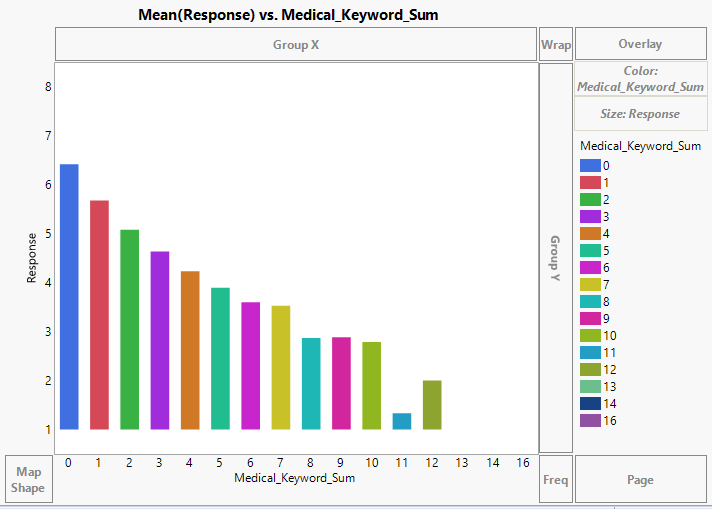
**Explore - Visualization of the data**

Data exploration is one of the important data preprocessing step. It is with the help of this step the accuracy of the classification of model can be taken to the next level.

Following are the findings developed based on the data.

***Finding#1*** – It is intuitive that persons with high Medical illness are more prone to high risk and so it is of higher risk to give insurance to such persons. The same has been observed from the data.

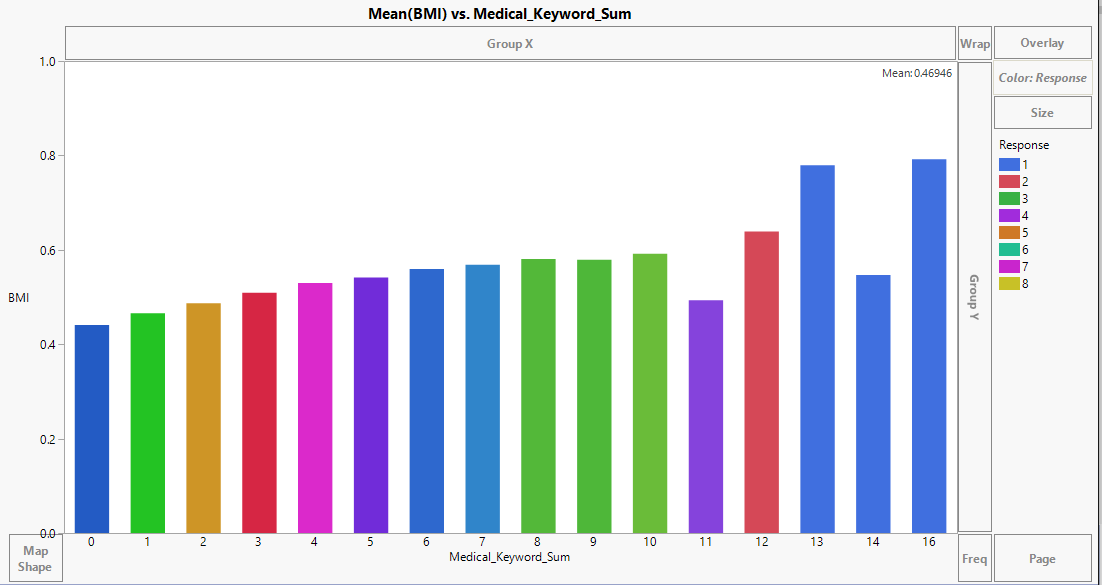
Following plot of Response Vs Medical\_Keyword\_Sum. Variable - ‘Medical\_Keyword\_Sum’ (explained in *feature engineering* section)



*Fig 1: Plot of ‘Response’ Vs ‘Medical\_Keyword\_Sum’*

***Finding#2*** – It is also intuitive that persons with high BMI are more prone to Medical Illness. The same has been observed from the data. In addition, it can also be observed that persons with low BMI are also highly prone to medical illness.

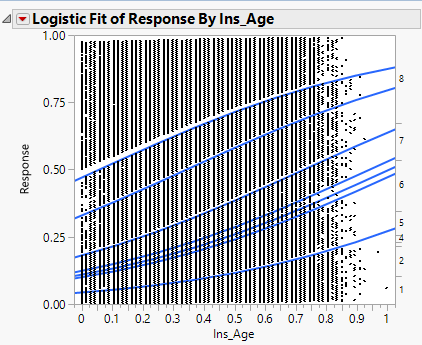
Following plot of BMI Vs Medical\_Keyword\_Sum. It can be observed that people with high BMI are having more medical illnesses when compared to people with lower BMI.



*Fig 2: Plot of ‘BMI’ Vs ‘Medical\_Keyword\_Sum’*

***Finding#3*** – It is a general observation that older people are more susceptible for higher risks. The same has been observed from the scatter plot between ‘Response’ and ‘Age’.

It is observed from the below plot that people with more age are more prone to risk when compared to people with lower Age.

****

*Fig 3: Plot of ‘Response’ Vs ‘Ins\_Age’*

**Modify - Data Modification**

Any dataset has to pass through this stage for successful prediction because all the datasets are prone to have missing values and outliers. They have to treated as per their impact on the target variable.

**Exploring Missing values**

Most of the missing values in the dataset are available in continuous variables. This may be attributed to the fact that the customers have a general tendency to answer application questions in which options are provided and they tend to ignore the ones in which they have to provide inputs on their own. Below are the missing variables in the dataset and hypothesized reasons as to why these may be prevalent

* **Medical History**

A significant amount (around 90%) of missing values has occurred for certain Medical History variables and all the missing data is for discrete variables only. People may not be willing to disclose their medical history details because they may fear of higher premium rates or possible rejection of life insurance due to prior illness. The variable under question could be the one that related to the number of years’ people have quit smoking/drugs and this question may not apply to all applicants. The question variable could also be related to the measure of blood pressure or sugar levels and people may not have the related information readily available.

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | N Missing | Percentage  Missing | Remarks |
| Medical\_History\_1 | 8889 | 14.96943467 | Discrete/Impute |
| Medical\_History\_10 | 58824 | 99.06198953 | Discrete/Ignore |
| Medical\_History\_15 | 44596 | 75.10146343 | Discrete/Ignore |
| Medical\_History\_24 | 55580 | 93.59896263 | Discrete/Ignore |
| Medical\_History\_32 | 58274 | 98.13576733 | Discrete/Ignore |

*Fig 4: Showing list of missing values for Medical History*

* **Family History**

People may not be willing to share the family history information to the insurer or they may not be aware of their family history.

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | N Missing | Percentage  Missing | Remarks |
| Family\_Hist\_2 | 28656 | 48.25786026 | Continuous/Ignore |
| Family\_Hist\_3 | 34241 | 57.66322561 | Continuous/Ignore |
| Family\_Hist\_4 | 19184 | 32.30663007 | Continuous/Ignore |
| Family\_Hist\_5 | 41811 | 70.41141106 | Continuous/Ignore |

*Fig 5: Showing list of missing values for family History*

* **Employment information**

People working in a hazardous working environment such as exposure to radioactive substances may not be willing to reveal their employment details. Some people may be unemployed for a certain period of time and they do not wish to provide their job details.

|  |  |  |  |
| --- | --- | --- | --- |
| Column Name | N Missing | Percentage | Remarks |
| Employment\_Info\_1 | 19 | 0.031996767 | Continuous/Impute |
| Employment\_Info\_4 | 6779 | 11.41610953 | Continuous/Impute |
| Employment\_Info\_6 | 10854 | 18.27857395 | Continuous/Impute |
|  |  |  |  |

*Fig 6: Showing list of missing values for Employment Information*

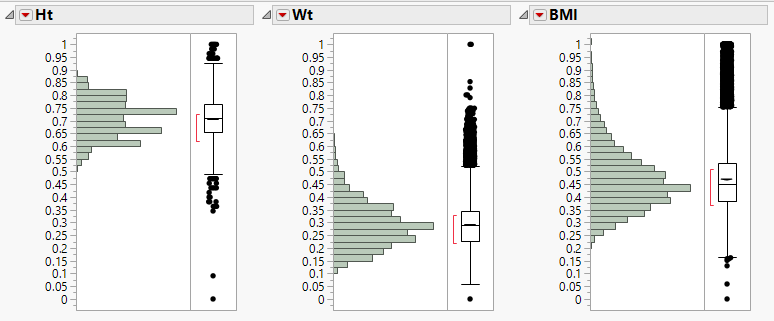
As not many values are missing in Medical\_History\_1, Employment\_Info\_1, Employment\_Info\_4 and Employment\_Info\_6; all those values are imputed and for the remaining variables lot of values are missing, so they are ignored.

**Exploring Outliers**

The following strategies are adopted to deal with outliers in the dataset.

* **Ignore outliers**

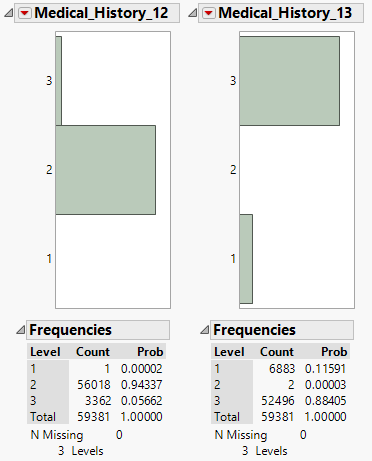
The outliers in the variables such as Age, Height and BMI are realistic because it's quite natural for people to fall apart from the normal distribution values. Hence, the outliers of these variables are ignored during processing.



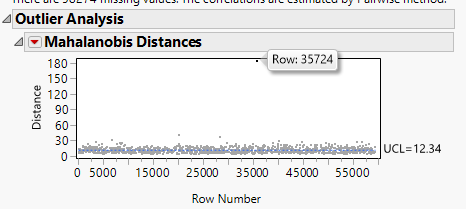
*Fig 7: Showing details of outliers which might be realistic*

* **Delete outliers**

The outliers in some variables may be because of typing errors or they could be fake data entries. So those outliers have been deleted.



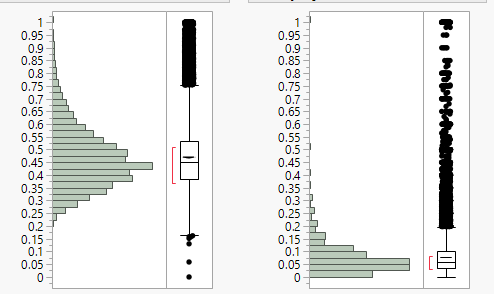
*Fig 8: Showing details of variables which might be a typing error*



*Fig 9: Showing details of variables with outliers for which is deleted*

* **Transforming outliers**

We have transformed some outliers which could really impact our modeling process. We have transformed them using the Mahalanobis transformation.



*Fig 10: Showing details of variables with outliers for which transformation is made*

**Feature Engineering**

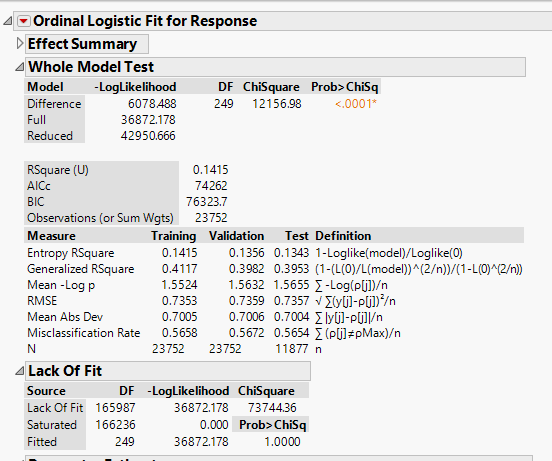
Addition of new features and removal of unwanted features would lead to better accuracy during the predictions. There are 48 ‘Medical Keyword’ variables which indicate the kind of the illness the person suffers. A new variable – ‘Medical\_Keyword\_Sum’ has been created which is the sum of all Medical Keyword variables in the corresponding record. Reduction of 48 variables to a single variable would reduce the modeling time.

**Model - Machine learning models**

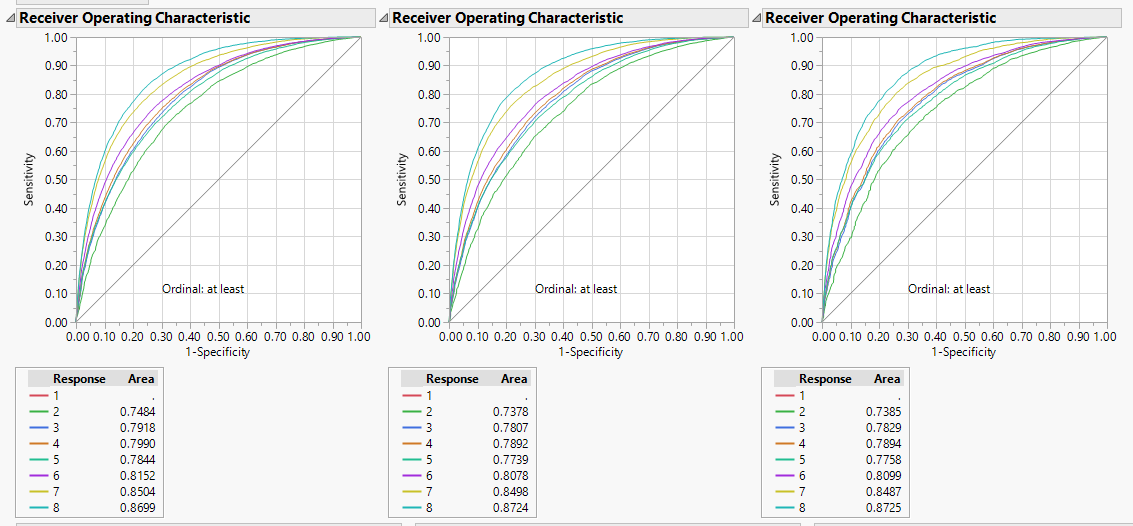
Three models are run to predict the risk associated with the insurance category. These models are ordinal Logistic Regression, Neural Network & Bootstrap Forest. Models which gives the most optimum results with respect to accuracy, the number of variables would be accepted. The number of variable plays a significant role as this will relate the time with money. We discarded the variables which have high missing values or huge outlier which is already discussed in the modify section of this report. The first model which we run is an ordinal logistic regression.

**Ordinal Logistic Regression:**

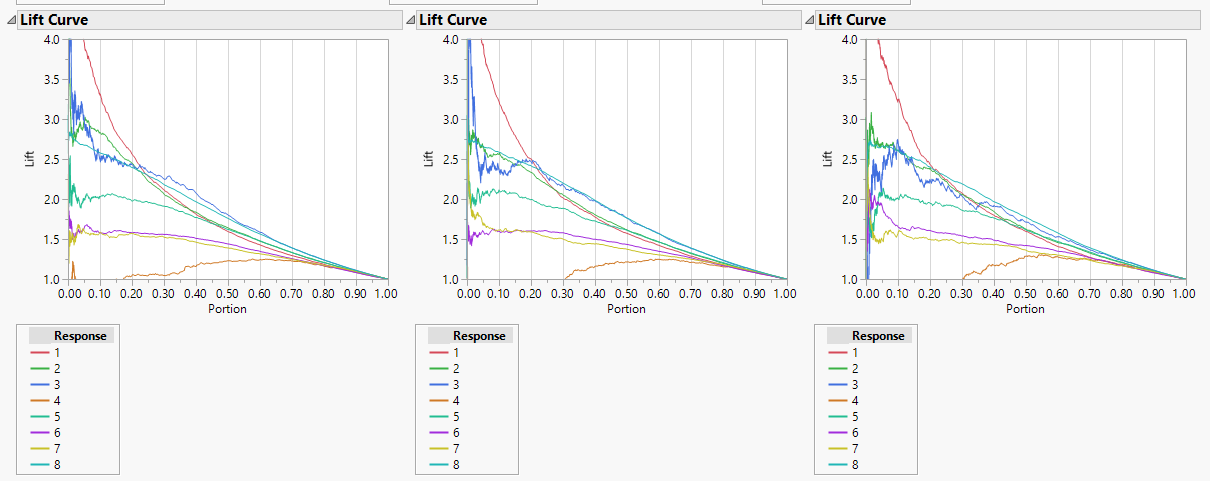
Ordinal logistic regression gives us the output which is shown below.



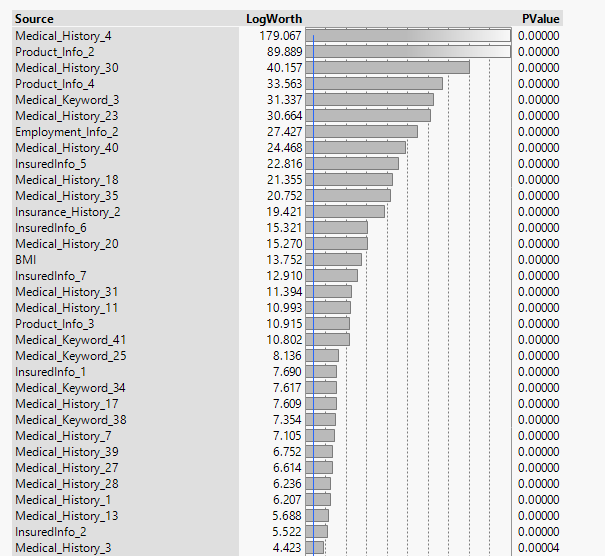
*Fig 11: Showing whole model test & summary*



*Fig 12: Showing ROC Curve & area values*



*Fig 13: Showing Lift Curves*

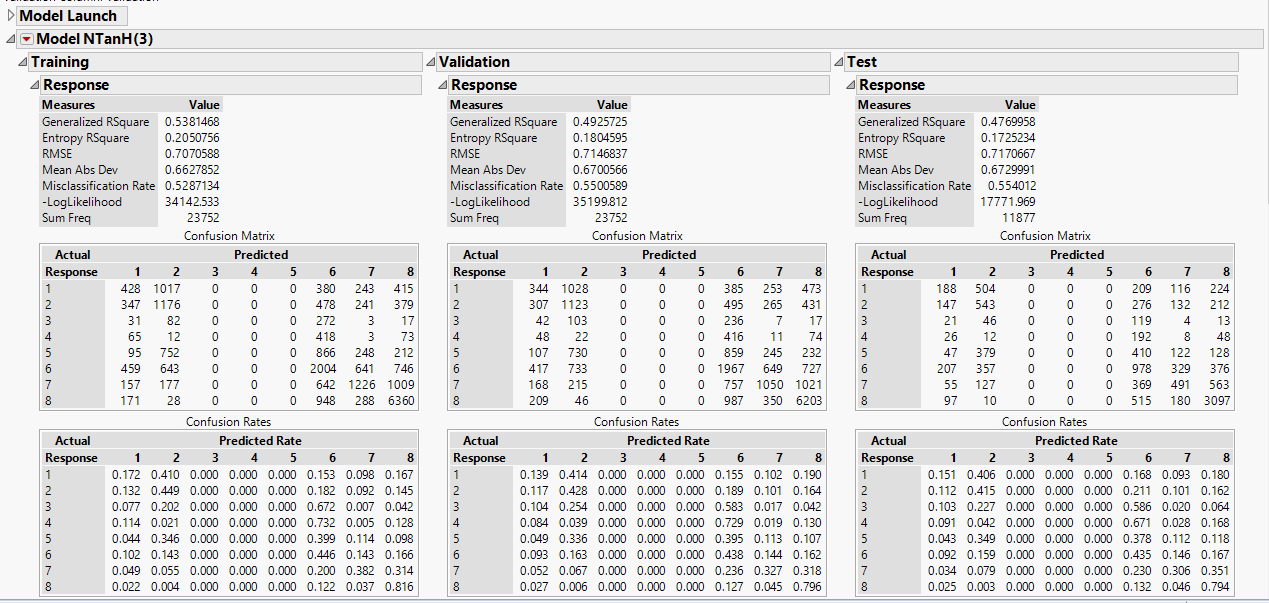


*Fig 14: Showing Log worth of the variables*

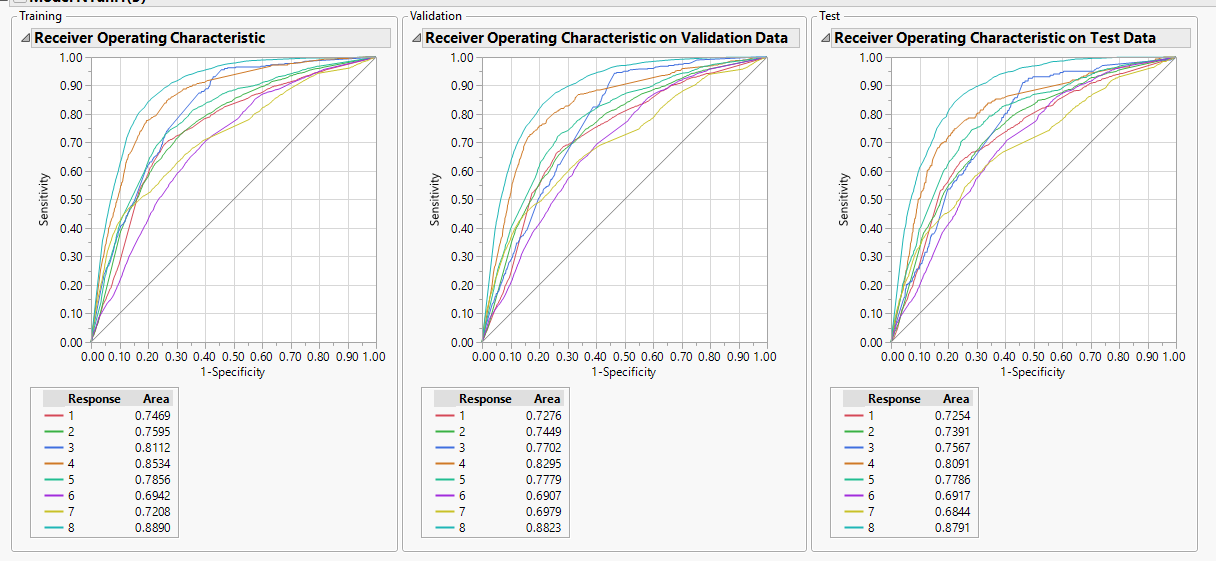
In ordinal logistic regression, from the above figures it can be inferred that R-Square is low for this model which is near to 0.1415 but the ROC curves gives very good results. ROC area is greater than 0.7 which implies the prediction is good. Significant predictors are reduced to half but still it has more than 30 predictors.

**Neural Network Model**

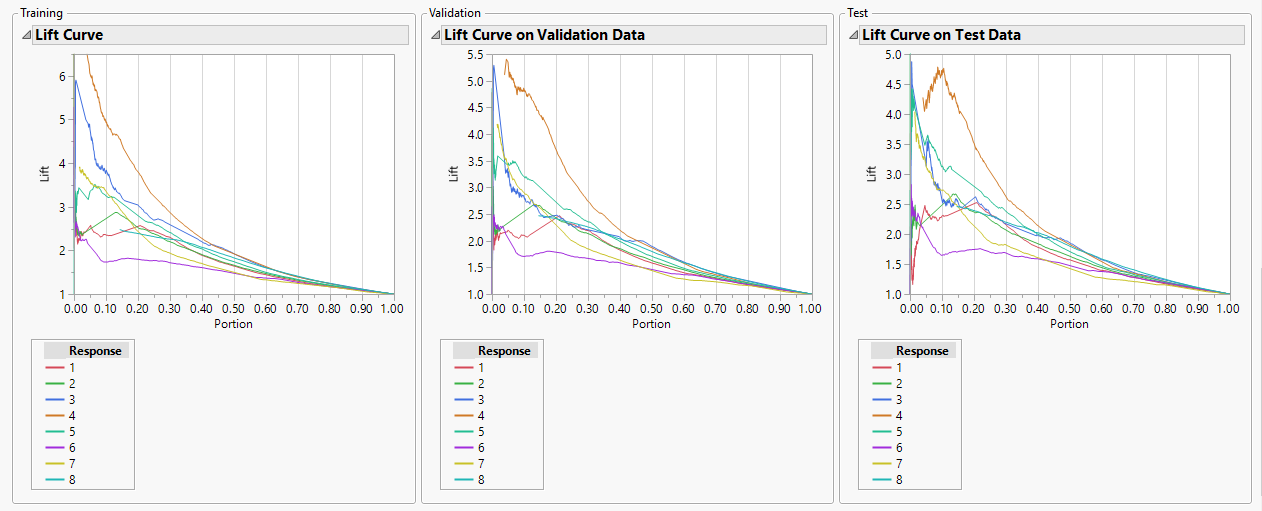
Neural Network gives us the output which is shown below



*Fig 15: Showing summary for Neural Network Model*



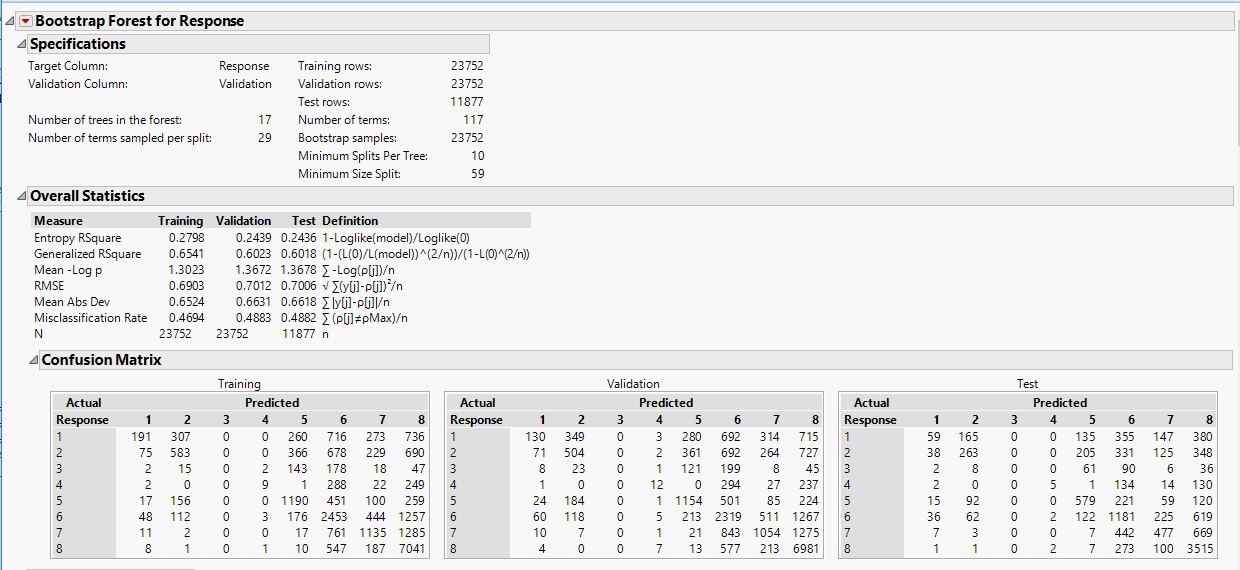
*Fig 16: Showing ROC Curve for Neural Network Model*



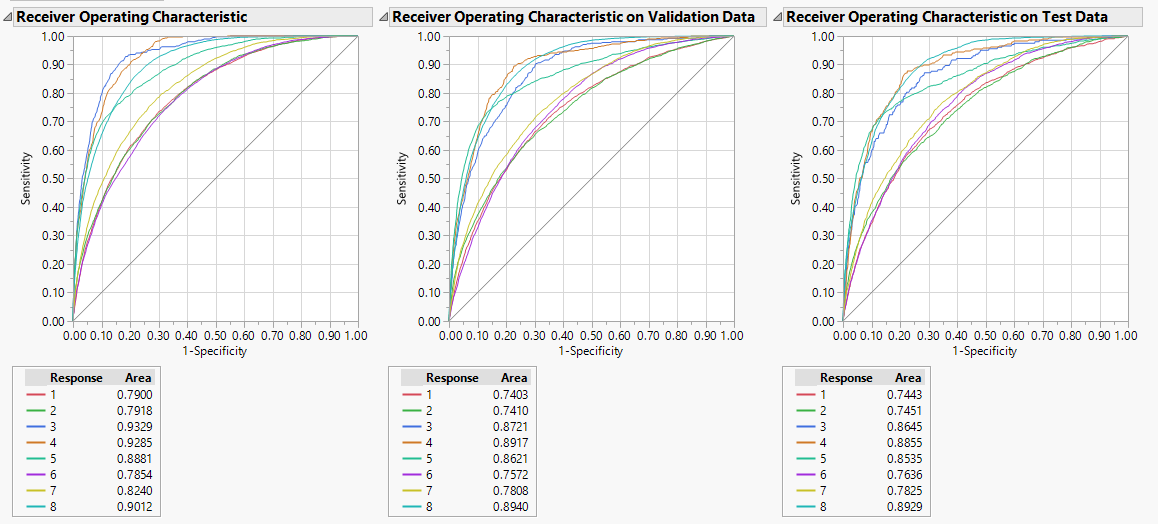
*Fig 17: Showing LIFT Curve for Neural Network Model*

In Neural Network, from the above figures we can infer that Rsquare is quite acceptable for this model and the ROC curves gives very good results. ROC area is greater than 0.7 which implies the prediction is good. Misclassification rate is also low. But as the Neural Network is very difficult to explain and every time it gives different Rsquare value another model need to be checked.

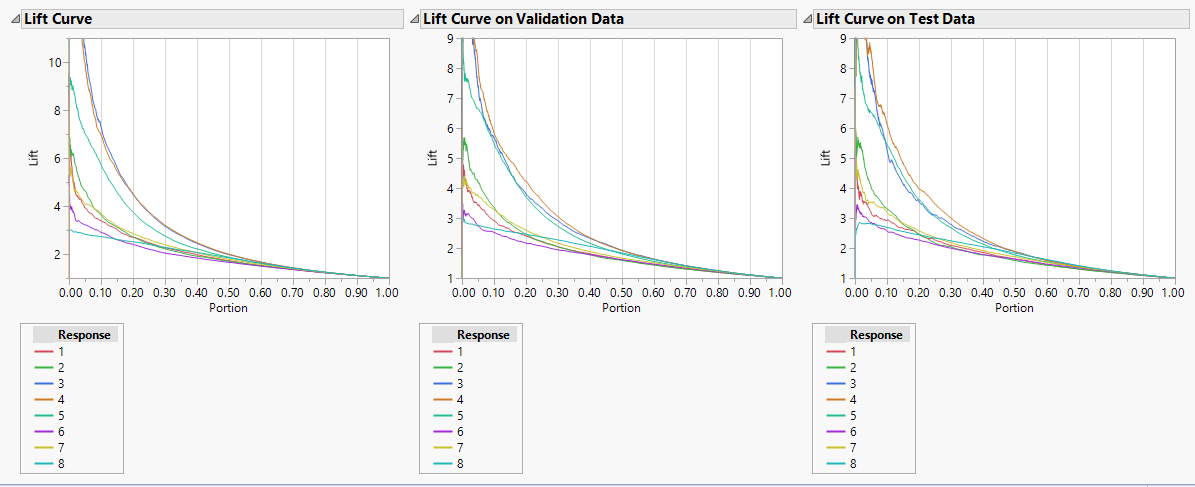
**Bootstrap Forest:**



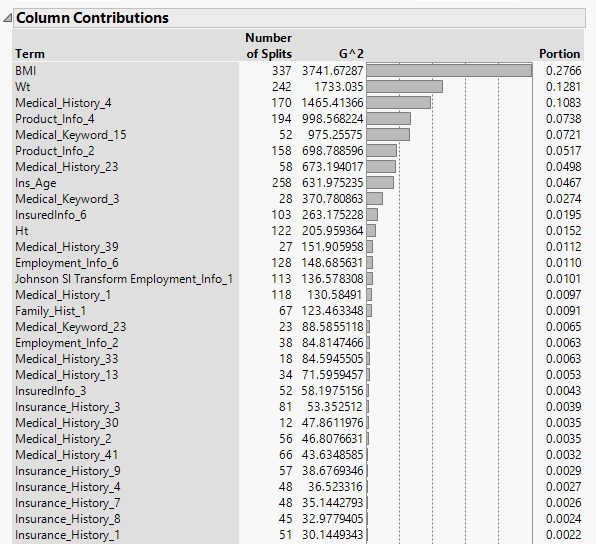
*Fig 18: Showing summary for Bootstrap Forest*



*Fig 19: Showing ROC curve for Bootstrap Forest*



*Fig 20: Showing Lift curve for Bootstrap Forest*

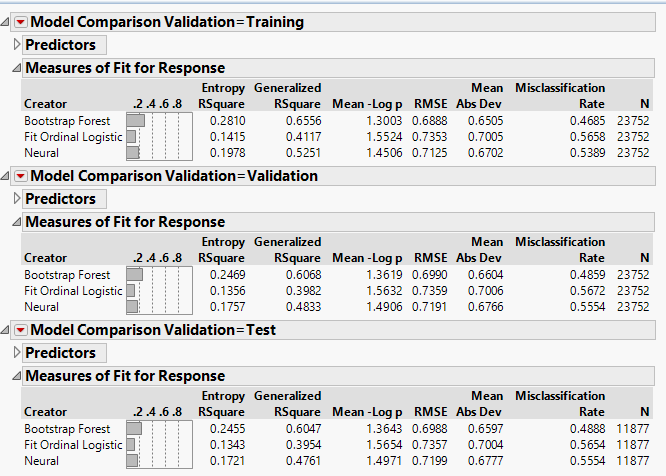


*Fig 21: Showing column contribution for Bootstrap Forest*

In Bootstrap forest, from the above figures we can infer that R-Square is quite ok for this model which is near to 0.281 but the ROC curves gives very good results. ROC area is greater than 0.7 which implies the prediction is good. Significant predictors are reduced to even less than one-fourth of the total variables.

Since all three models are giving similar results & one factor is dominating in each model so it is necessary to check the model comparison report.

**Model Comparison**



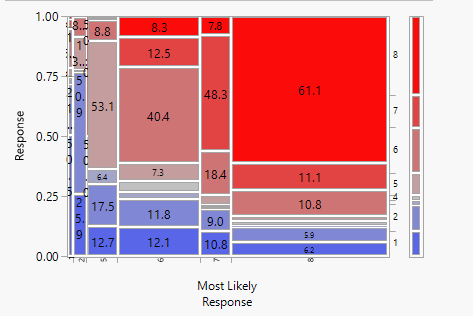
*Fig 22: Model Comparison Summary*

All the three model comparison summary are shown above. As R-Square value is less & it is the case of classification, other parameters are needed to check to finalize the model which we are going to accept. These include ROC curve & classification plots. ROC curve is showing good values in all the cases with best in the bootstrap forest. ROC curve plot & values are shown individually in models results. So, it is better to choose bootstrap forest model as the final model for classifying on the test dataset.

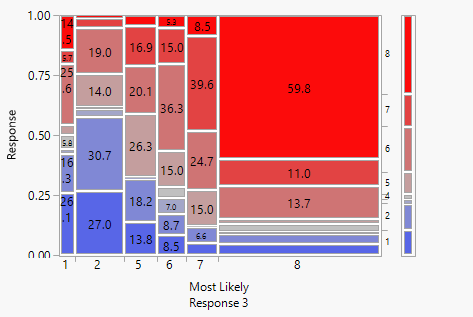
**Classification Rate:**

Following are the snapshots of misclassification matrix for all the 3 models that are executed on this dataset. Though it is difficult to analyze the misclassified elements because of large number of elements, it can give a big picture of how each model is performing.

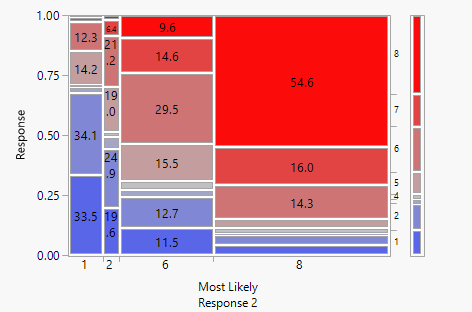
**Training Set:**



*Fig 23: Training set for Bootstrap Forest*

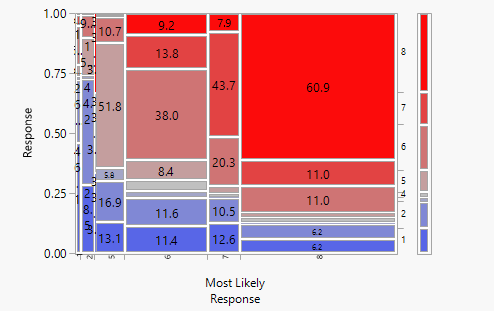


*Fig 24: Training set for Neural Network*

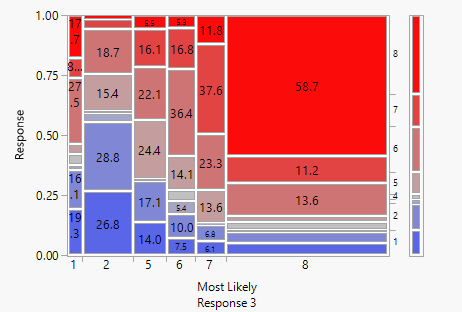


*Fig 25: Training set for Ordinal Logistic Regression*

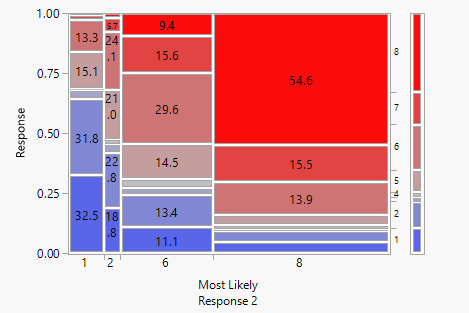
**Validation Set:**



*Fig 26: Validation set for Bootstrap Forest*

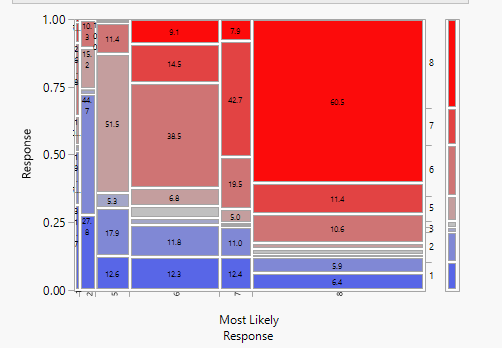


*Fig 27: Validation set for Neural Network*

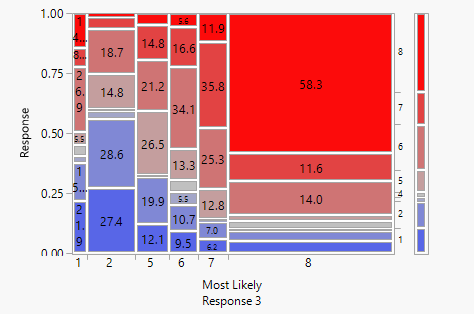


*Fig 28: Validation set for Ordinal Logistic Regression*

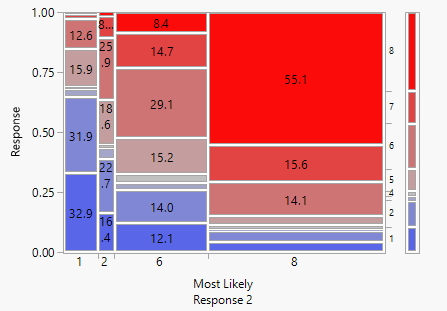
**Test Set:**



*Fig 29: Test Set for Bootstrap Forest*



*Fig 30: Test Set for Neural Network*



*Fig 31: Test Set for Ordinal Logistic Regression*

**Assess - Business Application**

Our model will genuinely solve several business needs within the insurance industry. For example:

* + - Save an insurmountable amount of time for both the company and the consumer
    - Reduce resources (and in turn, costs) required to acquire new customers
    - Attract significant market share of new insurance applicants given our model lends itself to the “instant gratification” society in which we now live
* Can be done at the leisure and convenience of the consumer, regardless of day and time
  + - Allows data to be collected in a centralized and concise manner, where it can be warehoused and referenced for years to come
* Ability for consumers to develop an online “profile” and keep track of previous quotes/discussions as it pertains to the information they provided
  + - Companies can standardize the ability to input certain parameters (i.e., drop down menus or multiple choice rather than input fields), significantly reducing the missing data that stifles their analysis abilities
* Allows for further inferences and conclusions to be drawn about the applicant
  + - Companies that employ the use of technology to benefit their processes may be seen as innovative and “cutting edge,” therefore attracting more market share by appealing to the younger demographic (who represent a significant amount of those seeking insurance for the first time)
* Potential to be an industry leader and pioneer, unlocking the market share that comes as a “first-mover”
  + - The company can satisfy the needs easily and quickly by deploying this model to first-time insurance buyers or “insurance shoppers,” where there is the potential to result in a significant amount of long-time or even life-long customers

**Conclusion**

Overall, the ease of use, utilization of technology, and forward-thinking innovation that would be evident by employing our model would have a positive effect on customer attraction and retention in the historically highly competitive insurance industry. By using a standard online model, the company would be able to harness big data in relation to their potential customer base, which could be applied and have significant impact on various business decisions for a vast amount of time. It is evident from the above mentioned reasons coupled with potential cost savings, the implementation of this model would be a noteworthy win-win for both the corporation and prospective customers.