PRUDENTIAL LIFE INSURANCE RISK ASSESSMENT - MILESTONE REPORT

IDS 575 – Business Analytic Statistics - Final Project

Abstract

To help the insurance companies in accessing risk of their customers in an automated and systematic way to save on time and cost

ASHOK BHATRAJU - 670248723

NIKITA BAWANE - 661069000

RITU GANGWAL - 670646774

PRUDENTIAL LIFE INSURANCE RISK ASSESSMENT - MILESTONE REPORT

Introduction:

Prudential Life Insurance uses individual customer's data to assess risk in providing insurance. The traditional methods include collecting medical history, family records and insurance history among many other data points, the process usually takes 30 days. Our goal is to automate the process so that risk assessment would be quicker, cost effective and more accurate.

Our approach includes:

- Data preparation, cleaning and customization
- Understanding data by performing exploratory data analysis
- Building predictive models to Identify key variables which contribute in predicting response variable (determining level of risk, High 1 or Low 0)
- Model analysis, selection and reporting results

The results of this modelling would help the insurance companies to make well informed decisions.

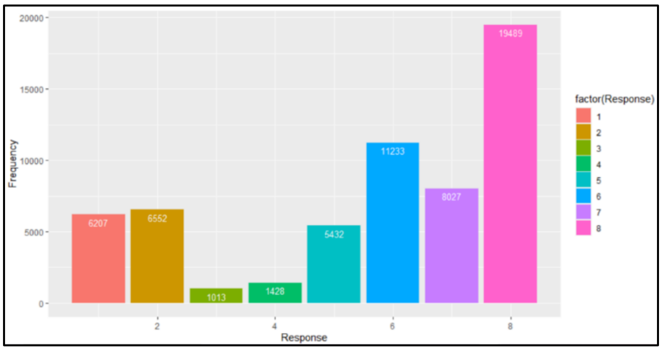
Variable	Description	Variable Treatment Summary
ID	A unique identifier associated with an application	This variable is not important in model development as it is just a reference variable.
Product_info_1-7	A set of normalized variables relating to the product applied	All these variables are treated firstly through missing values, replacing these missing values and then applying random forest method to get the important variables.
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	
Family_Hist_1-5	A set of normalized variables relating to the family history of the applicant	
Insuredinfo_1-7	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	PCA method is applied for variable reduction along with all information restored.
Medical_History_1-41	A set of normalized variables relating to the medical history of the applicant	Created new variable as "Medical History Sum" which is basically the sum of all 3 to 41 variables to restore all the information. This new variable will represent medical history of an applicant in total. Medical History 1 and 2 are left as it is as they have high values.
Medical_keyword_1-48	A set of dummy variables relating to the presence of/ absence of a medical keyword being associated with the application	Created new variable as "Medical Keyword" which is basically the sum of all dummy variables to restore all the information. This new variable will represent total no. of medical keywords present for an applicant

Target Variable	Description	Variable Treatment Summary
Response	This is the target variable, an ordinal variable relating to the final decision associated with an application	Converted this into binary variable - 0 for ranks 1 to 4 as "Low Risk Applicants" and 1 for ranks 5 to 8 as "High Risk Applicants"

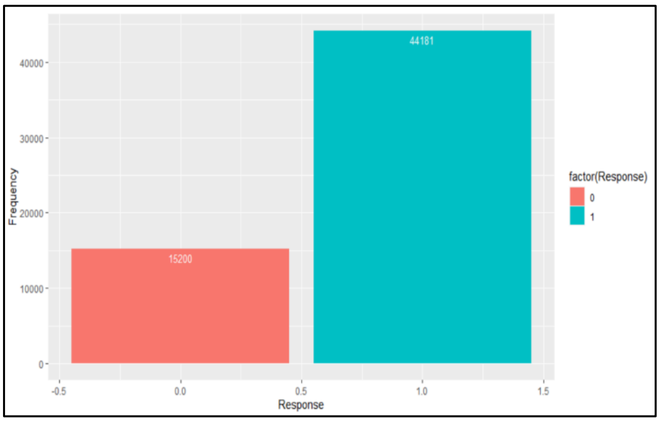
Data Preparation and Exploration:

The data set is pre-separated among training and test sample in the ratio of 3:1 and have a random sampling being done. The train data set is a transactional data consists of 59,381 customers and 126 variables as predictors. The test dataset contains the same variables for another set of 19,765 customers. (Source: https://www.kaggle.com/c/prudential-life-insurance-assessment/overview)

Converting multinomial response variables into binomial



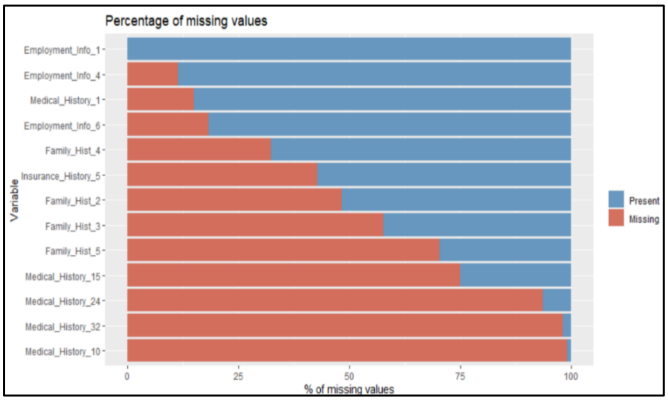
Initially our training data set consists of Response variable ranging from 1-8. This range helps the firm in identifying the chances of insurance claim by the customer. We have converted it to a binomial variable. The frequency distribution plot of the Response variable is shown on the left.



We converted the multinomial response variables into binary (0,1). 0 represents “Low Risk Applicant” and 1 represents “High Risk Applicant”. Preference should be given to 1.

Response Variable	Frequency
0	15200 (25%)
1	44181 (75%)

Missing Values Treatment



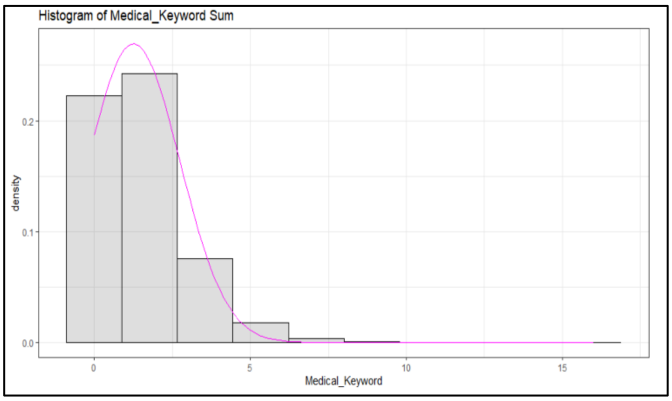
The missing value graph depicts that there 13 variables which have missing values in them. The variables with more than 70% missing values are removed during model development. The missing values for other variables have been replaced by their median values after analysing their normal distribution curves.

Summary of missing value treatment is shown in the table below:

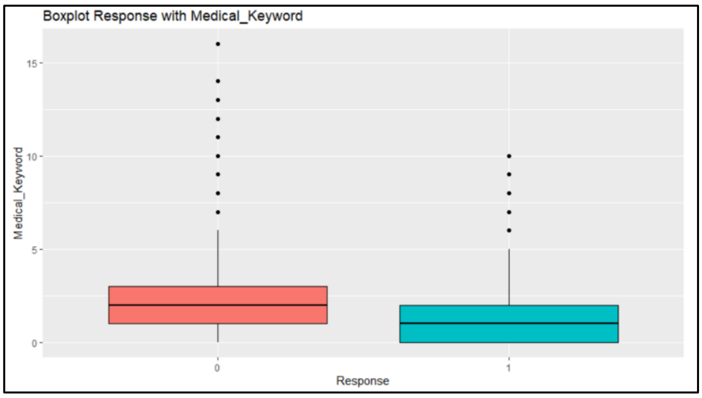
Missing value percentage	Variable List	Action taken
More than 70% values missing	Family_Hist_5, Medical_History_15, Medical_History_24, Medical_History_32, Medical_History_10	Removed the variables
Less than 70%	Employment_Info_1, Employment_Info_4, Medical_History_1, Employment_Info_6, Family_Hist_4, Insurance_History_5, Family_Hist_2, Family_Hist_3	Replaced missing values with Median

New Variables Introduced:

a. MEDICAL_KEYWORD



Histogram of derived variable Medical_Keyword sum

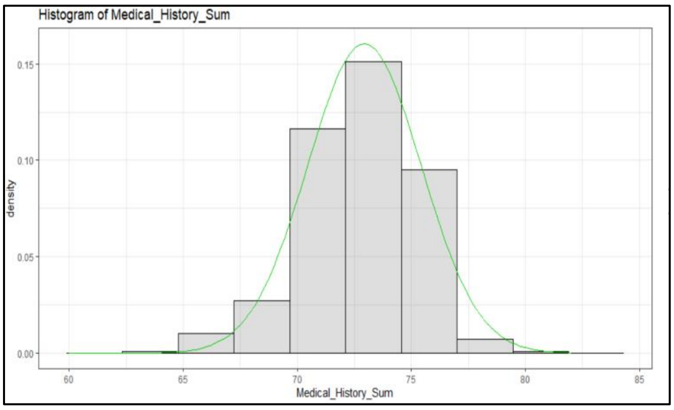


Boxplot of derived variable Medical_Keyword sum

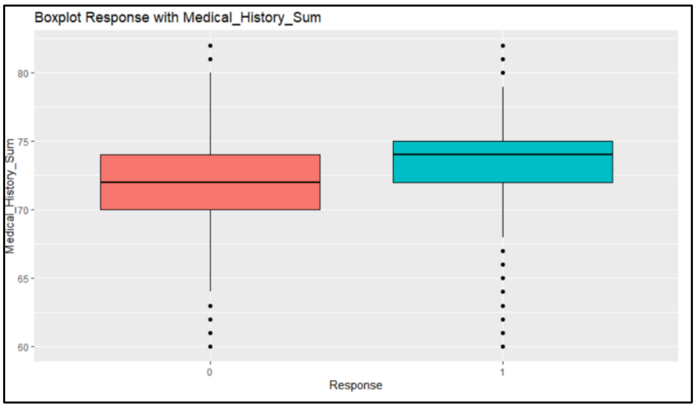
There are 48 'Medical_Keyword_' variables in the dataset. These variables mostly consist of 0's and 1's. Individually, these variables have very less predictive power. Hence, they are combined, by taking the *sum*, to form a single variable 'Medical_Keyword'.

The graph on the left shows that the derived variable in not normally distributed. The maximum number of values fall in the range of 0 to 5. The graph on the right shows that Medical_Keyword distribution for each response variable (0 and 1) and we can conclude that the new variable has fair predictive power.

b. MEDICAL_HISTORY_SUM



Histogram of derived variable Medical_Keyword sum



Boxplot of derived variable Medical_Keyword sum

Medical history is spread across 41 different variables. We combine variables from 3 to 41 to form a single variable Medical History Sum. Initial 41 variables were then deleted from data.

The graph on the left shows that the derived variable in normally distributed. The graph on the right shows that Medical_History_Sum distribution for each response variable (0 and 1) and conclude that the new variable has good predictive power.

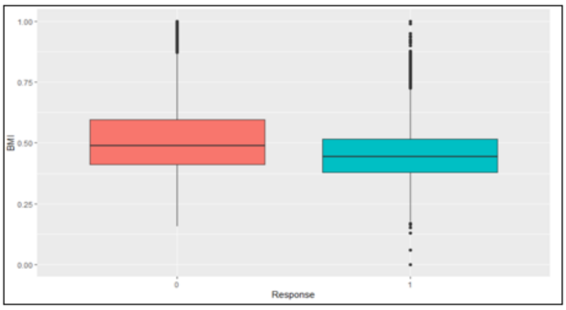
This method of summing has led to large variable reduction with all the information intact. More the value of new derived variables, more should be chances of risk as it refers to severe medical conditions in the past and in present.

Data Exploration for Continuous Variables:

In order to understand the relation between continuous independent variables i.e. Height, Weight, BMI, Product Info2 and Age; and response variables, we have used box plots and correlation table analyse the relationship. These variables are treated independently and examined closely with that of the target variable Response.

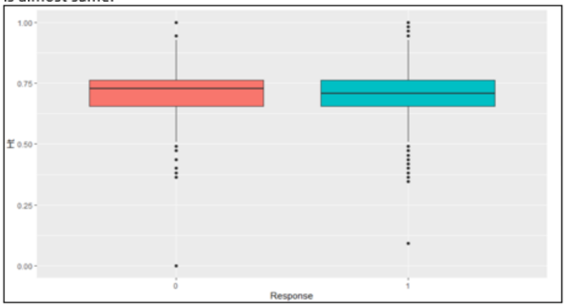
BMI vs Response

BMI has fair predictive power as the medians for the response variable 0 and 1 are slightly varying.



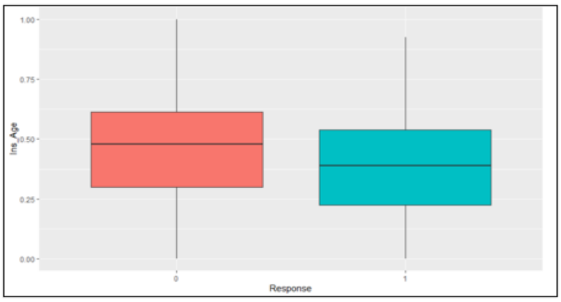
Ht vs Response

Ht has very less prediction power. The medians of Ht for the response variable is almost same.



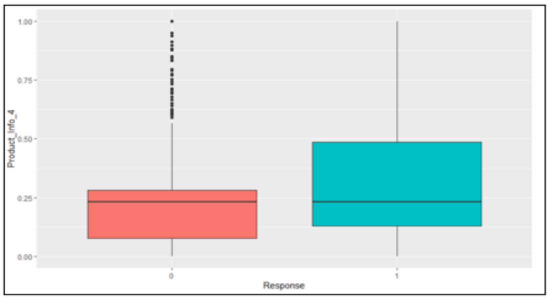
Ins_Age vs Response

Ins_Age has fair predictive power.



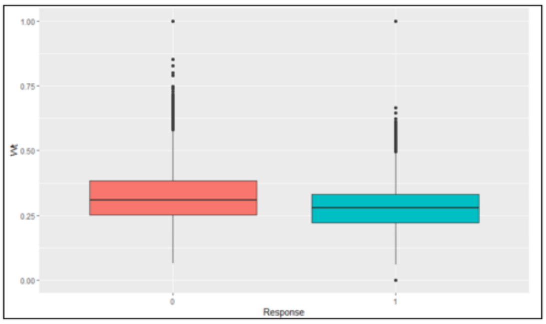
Product_Info_4 vs Response

Product_Info_4 has very less predictive power.



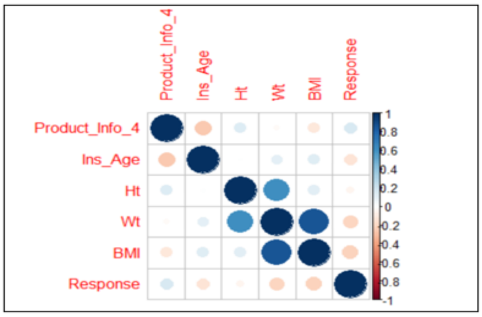
Wt vs Response

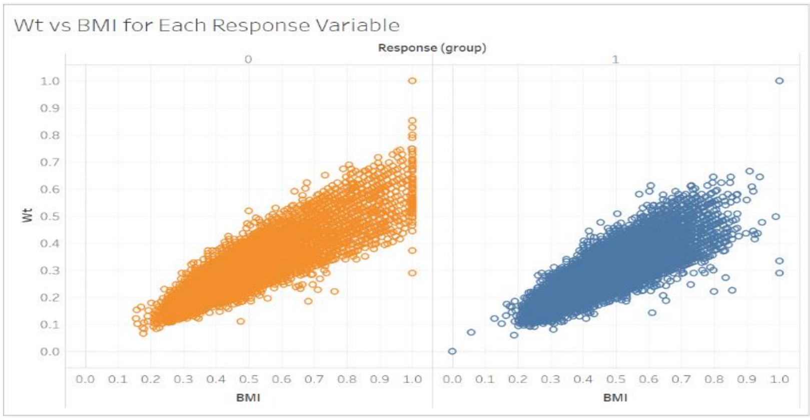
Wt has less predictive power.



Correlation Between Important Continuous Variables

BMI and Wt are highly correlated (<value>).





We can infer from the correlation plot that BMI and Wt are highly correlated. It is an obvious observation as the value of BMI is dependent on individual's height and weight. Other continuous variables are not correlated with our output variable – Response.

__Principal Component Analysis – Dimensionality Reduction:__

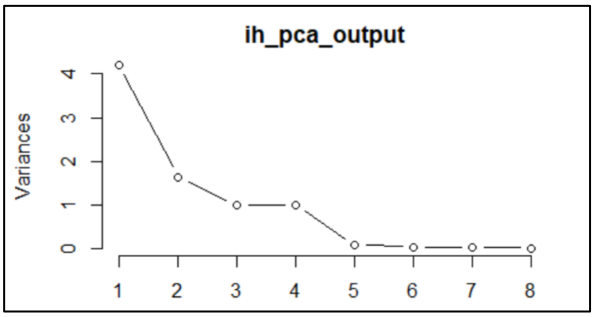
We performed Principal Component Analysis on different groups of variables (as mentioned in the table below) in order to reduce the dimension of the feature space. We found that PCA was most effective for 'Insurance_History' variables as its 4 principal components retained the maximum information of the 9 variables. Hence, the 4 principal components are added to the dataset and original 9 Insurance History variables are removed.

In the graph below, we can select which features to include by determining whether the addition of another feature results in a significant drop in variance relative to the previous feature and retaining features till that point.

Variables	PCA Output Analysis	PCA Outcome
Product_Info (1-7)	By selecting 6 Principal Components(PC) out of 7 PC , we can only preserve 89.46% of the total variance of the Employment_Info data	Since the number of feature are not reduced significantly, PCA is not an effective method for feature reduction for Product_Info
Employment_Info (1-6)	By selecting 5 Principal Components(PC) out of 6 PC , we can preserve 96.10% of the total variance of the Employment_Info data	Since the number of feature are not reduced significantly, PCA is not an effective method for feature reduction for Employment_Info
InsuredInfo (1-7)	By selecting 6 Principal Components out of 7 PC , we can preserve 93.02% of the total variance of the Insured_Info data	Since the number of feature are not reduced significantly, PCA is not an effective method for feature reduction for InsuredInfo
Insurance_History (1-9)	By selecting 4 Principal Components out of 8 PC, we can preserve 98.06% of the total variance of the Insured_History data	First 4 Principal Components have the highest variance for Insurave_History variables, hence these components will be added in the dataset

__PCA for Insurance History__

```
Importance of components:
                          PC1    PC2    PC3
Standard deviation     2.0517 1.2817 0.9996
Proportion of Variance 0.5262 0.2053 0.1249
Cumulative Proportion  0.5262 0.7315 0.8564
                          PC4     PC5     PC6
Standard deviation     0.9966 0.27816 0.18710
Proportion of Variance 0.1241 0.00967 0.00438
Cumulative Proportion  0.9806 0.99027 0.99464
                           PC7    PC8
Standard deviation     0.17334 0.1132
Proportion of Variance 0.00376 0.0016
Cumulative Proportion  0.99840 1.0000
```



__Correlation among other variables:__

The variables which had correlation value more than or equal to 0.7 were removed (as shown in the output below). Also, variable 'Id' is removed as it does not add to prediction power of the model. The total number of variable remaining is 32.

```
All correlations <= 0.7
[1] "Wt"                "InsuredInfo_6"
[3] "Employment_Info_3" "Employment_Info_5"
```

__Random Forest for Variable Importance:__

After data exploration and feature engineering, we run Random Forest to determine the top 25 important variables (from Variable Importance). This is performed in order to get only those variables that play significant role in our modelling. As PCA and summation method can't be applied to all buckets of variables, this was an important step to recognize essential variables from family history, insured info, employment info and product info.

Baseline Models:

After all data exploration and feature engineering, we have all our training instances with 26 variables left. We have then divided our training data into training and validation data in 70:30 ratio. We have trained our data by the below two models of majority selection:

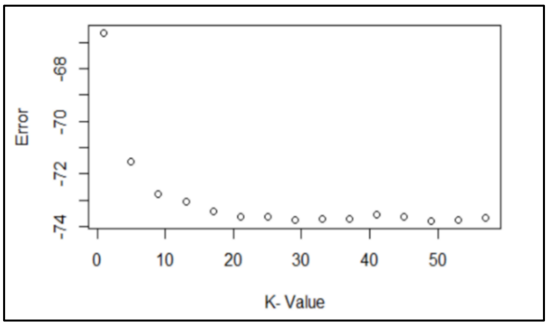
1. ***Naïve Bayes*** – Baseline model as it selects the majority class of the whole data and returns the output

```
> confusionMatrix(Predict_val, pdval$Response)
Confusion Matrix and Statistics

           Reference
Prediction  Low Risk High Risk
  Low Risk      1014       754
  High Risk     3596     12450

               Accuracy : 0.7558
                 95% CI : (0.7494, 0.7621)
    No Information Rate : 0.7412
    P-Value [Acc > NIR] : 3.988e-06

                  Kappa : 0.2037

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.21996
            Specificity : 0.94290
         Pos Pred Value : 0.57353
         Neg Pred Value : 0.77589
             Prevalence : 0.25879
         Detection Rate : 0.05692
   Detection Prevalence : 0.09925
      Balanced Accuracy : 0.58143

       'Positive' Class : Low Risk
```

Naive Bayes is a Supervised algorithm based on the Bayes Theorem that is used to solve classification problems by following a probabilistic approach. It is based on the idea that the predictor variables in a Machine Learning model are independent of each other.
In our model, we have already removed all the correlated variables, hence we can apply this model.
Accuracy of this model = 75.58% but less sensitivity of 21.99%.

2. ***KNN Method*** – This is a further improvement over Naïve Bayes as it returns majority class of the k-nearest neighbours. We have developed models with different values of K and found that the accuracy is maximum at K = 49 as seen from the below output. The first image shows accuracy at various values of K and second image depicts that error is least at K = 49.

```
1 = 67.6322
5 = 72.53284
9 = 73.79028
13 = 74.06534
17 = 74.42461
21 = 74.65477
25 = 74.63231
29 = 74.74458
33 = 74.70529
37 = 74.72774
41 = 74.57056
45 = 74.63793
49 = 74.78388
53 = 74.76142
57 = 74.68283
```



Milestone Summary and Timeline: