Capstone project-1: Prudential Life Insurance Full Report

Gautham Gowda

Contents

Problem statement- why is it useful to answer the question

Insurance policies give a sense of security to the policy holder in case of uneventful things in life such as death, disability. At the same time the policy issuer have to make sure they do not end up losing money if there are more claims paid out than the premiums collected for insuring people. This study/ model will use various data analysis techniques to predict the risk of insuring each customer based on risk factors such as age, weight, height, BMI, sex, medical history, employment history etc. The analysis also will address machine learning techniques to check the goodness of fit between the independent variables and the target variable.

Clients/ Intended audience

The study is of interest to any companies that offer insurance or loans to customers based on customer profiles to predict risk. The study already has classified the customers based on risk and this study works as a sanity check to improve model efficiency. Therefore, it can be applied to a variety of predictive analytics cases.

Even though this exercise involves term life insurance case, many other companies can benefit from this analysis, such as banks that offer other insurances, loans to consumers, health insurance market place and can be extended to other types such as home or auto insurances.
The analysis helps them to decide what types of customers are more risky to insure and come up with a model to predict insurance premiums based on risk.

Dataset used for the investigation

The dataset for this analysis comes from Prudential Life Insurance and attached is the link to the dataset.
https://www.kaggle.com/c/prudential-life-insurance-assessment/overview.

Since the data consists of private and sensitive information, many of the data labels (columns) are coded to make it ambiguous. For example, there are a list of variables containing applicants' health information. But instead of listing the actual variable label such as 'High blood pressure', it may be listed as 'Medical_History_1'. Similarly, a salary of the applicant can be listed as 'Employment_History_1'. After the analysis we will only know if which coded variables affect the analysis. Prudential insurance company will take the analysis and match these coded variables to the actual variables affecting the results.

However, some variables such as age, BMI, Wt, Ht are provided without masking them. But these variables are normalized to a scale between 0 and 1.

The following table summarizes all the variables:

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

The response variable is an ordinal variable categorized 1 thru 8 and relates to the final decision associated with the application. Again, we will not know exactly what each of these categories mean.

Data cleaning and wrangling

- Data cleaning: The data is taken from Kaggle competitions and is provided by Prudential Life Insurance. The dataset is pretty clean and the variables are already arranged as columns. The variable names are coded to maintain privacy of the information. All the continuous variables are normalized.
- Missing values: There are variables with missing values. Some of these variables are continuous and some are discrete and categorical. Used the pandas data-frame method to fill the missing values in the columns:

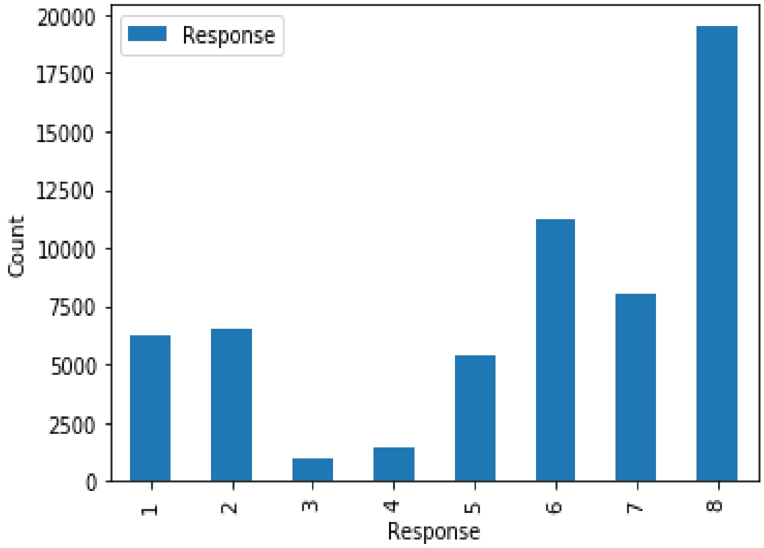
```
data = df.fillna(method='ffill').fillna(method='bfill')
```

  This method to replace all missing values with forward fill and backward fill data.

- Outliers: There is significant outliers present in some variables. Variables ‘Age’, ‘BMI’ and other continuous variables have some outliers. After plotting the response vs the variable of interest as bar chart, the effect of outliers seemed insignificant. Also, the outliers appear to be good data points rather than outliers. Tried choosing data within 15-85 percentiles to reduce the outliers but the effect one the data is minimal. Another drawback with eliminating outliers is loss of nearly 8000 observations of data. Therefore, even though, removing the outliers helps with data visualization, all the data is kept as is for the analysis.

Data visualization

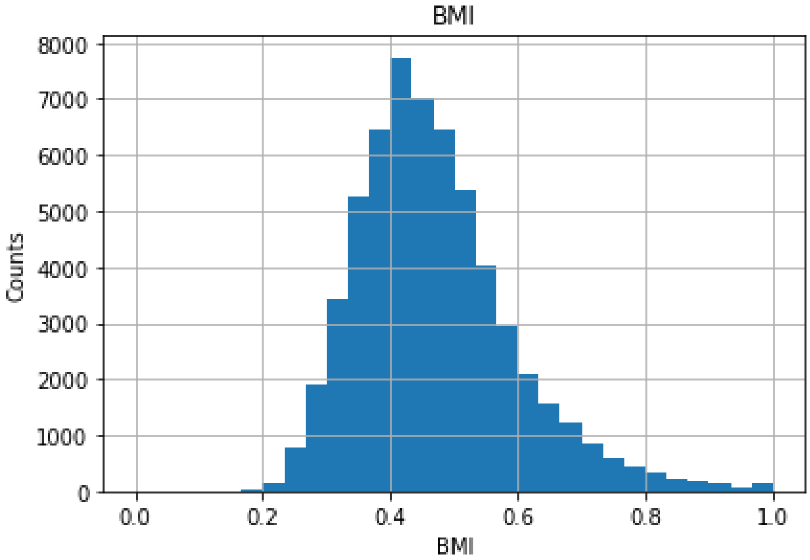
Following plots provide the visualization of the data for analysis. The continuous variables are plotted in histograms to check for normality. Other plots check the effect of some of the independent variables on the 'Response' variables grouped by the eight response categories.



Response	
1	6207
2	6552
3	1013
4	1428
5	5432
6	11233
7	8027
8	19489

Fig-1: Response variable counts for the eight categories

The above figure demonstrates the response categories for the eight decisions taken on the applications.



	BMI
count	59381.000000
mean	0.469462
std	0.122213
min	0.000000
25%	0.385517
50%	0.451349
75%	0.532858
max	1.000000

Fig-2: BMI distribution and summary statistics

The BMI histogram is shown here as an example to continuous variables with normal distribution. Other variables such as age, height, weight, and some employment info variables (most likely salary info) show normal distribution.

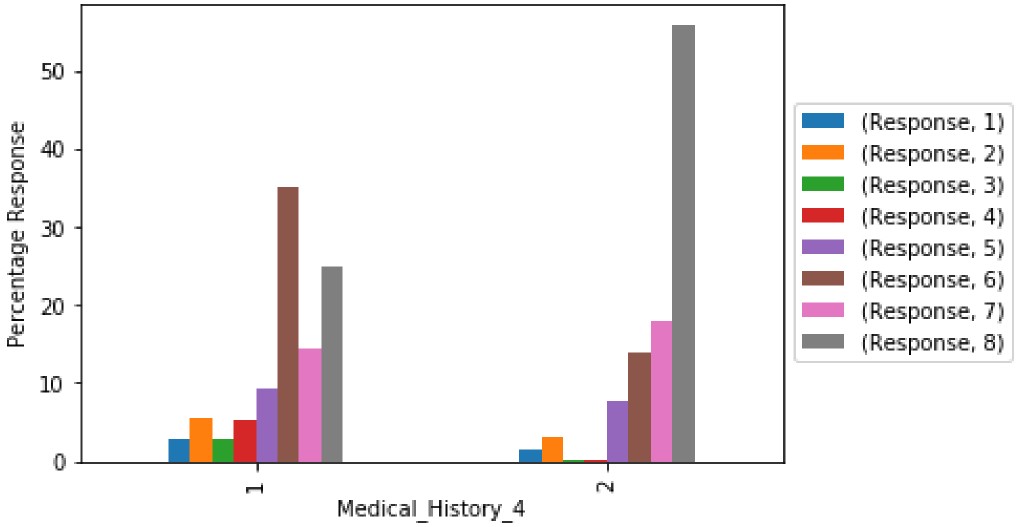


Fig-3: Categorical variable examples Medical_History percentages of Response

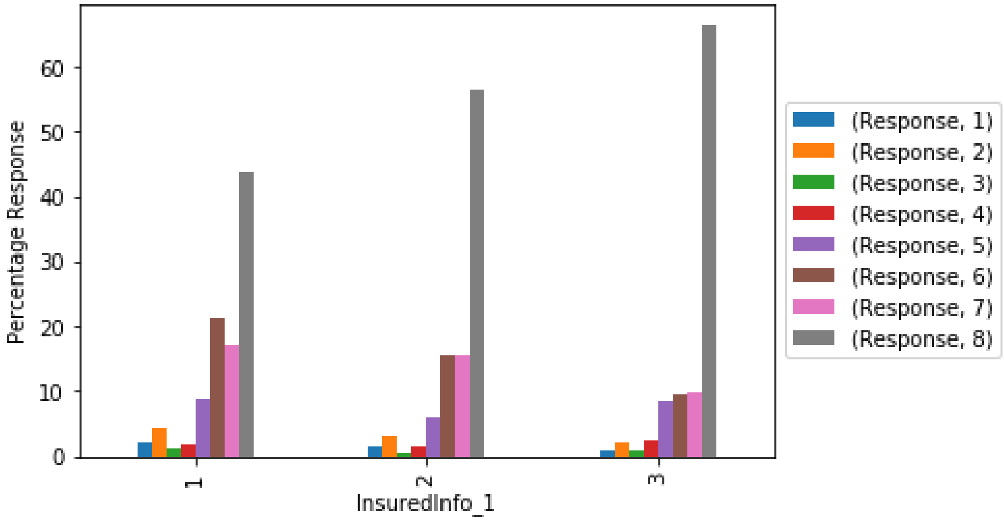


Fig-3: Categorical variable examples InsuredInfo_1 as percentages of Response

The above two figures represent the data distribution for categorical variables. Most of the variables are categorical. Again, as explained earlier, with labeling ambiguity it is hard to classify each categories. However, the effect on the eight categories as a percentage is shown in the above two plots.

Exploratory Data Analysis (EDA)

Are there variables that are particularly significant in terms of explaining the answer to your project question?

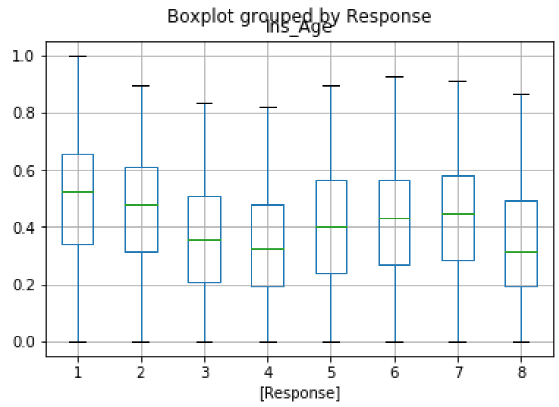
The ‘Response’ variable is the final decision associated with the insurance application. From the problem definition we can think of certain variables that can affect an insurance application. Some of the most significant ones are: BMI, age of the applicant, medical history of the applicant, employment status and family history of the applicant.

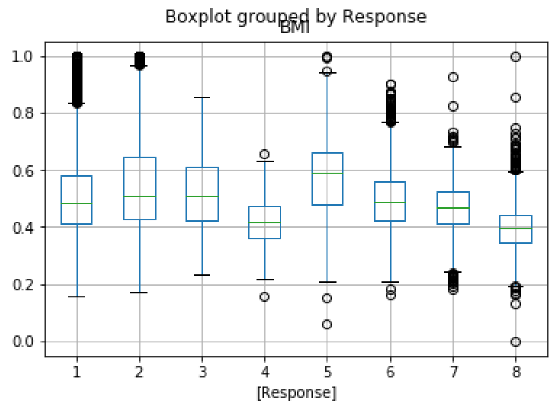
Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

Yes some of the independent variables have strong correlation between them. For example, BMI of an applicant is calculated with a person’s height and weight and we see a strong correlation between them using correlation coefficients between them. Checked also the correlation between the ‘Response’ variable and some of the independent variables.

What are the most appropriate tests to use to analyze these relationships?

- Visual analysis using the box plots to see the response variable vs the independent variables give us information about the relationships. Below we have two plots with independent variables Age and BMI of the applicants grouped by dependent variable ‘Response’.





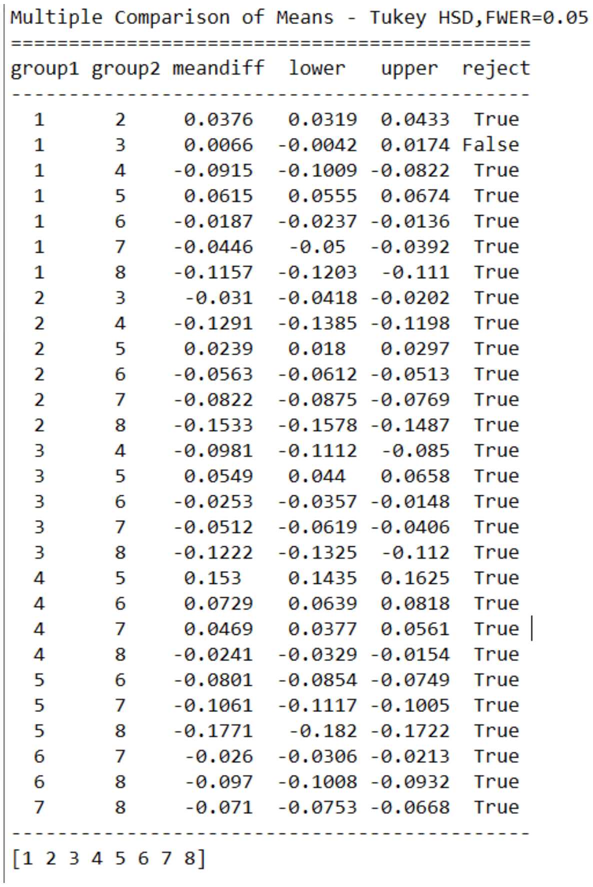
- The correlation matrix to check dependency between variables is done to see the effect of some of the variables on the dependent variable. Ran a correlation matrix to check some of the independent variables effect on the response variable. As summarized in the table below, the person’s age, BMI, weight are all inversely correlated to the response. As expected, with increase in age, weight, BMI of applicant, the approval goes down.

	Ins_Age	**BMI**	**Ht**	**Wt**	**Response**
Ins_Age	1.000000	0.137076	0.008419	0.110366	-0.209610
BMI	0.137076	1.000000	0.123125	0.854083	-0.381601
Ht	0.008419	0.123125	1.000000	0.610425	-0.093576
Wt	0.110366	0.854083	0.610425	1.000000	-0.351395
Response	-0.209610	-0.381601	-0.093576	-0.351395	1.000000

- Significance tests to check for p-value to see if the data is significant: The dataset has a total of 59300 data points. The sample size is large enough to be statistically significant. Also, conducted analysis of variance for f-stats and probability value. The null hypothesis (H0) is that the data is not statistically significant. Analysis of some of the important variables indicate the p-value = 0 which is less than alpha=0.05 and therefore reject the null hypothesis.

  Also, did multiple comparisons of means between the response categories vs the independent variables such as applicant age, BMI etc. Since P-value = 0 we can reject the null hypothesis that difference in means between categories of Response = 0.

  The data below shows the comparison of means for 'BMI'. As we can see, the mean difference is significant at alpha = 0.05 and therefore we can reject null hypothesis that there is so difference between the means of the response categories. Therefore, the data is significant.



```
Multiple Comparison of Means - Tukey HSD,FWER=0.05
=============================================
group1 group2 meandiff  lower   upper  reject
---------------------------------------------
  1      2     0.0376   0.0319  0.0433  True
  1      3     0.0066  -0.0042  0.0174 False
  1      4    -0.0915  -0.1009 -0.0822  True
  1      5     0.0615   0.0555  0.0674  True
  1      6    -0.0187  -0.0237 -0.0136  True
  1      7    -0.0446   -0.05  -0.0392  True
  1      8    -0.1157  -0.1203  -0.111  True
  2      3     -0.031  -0.0418 -0.0202  True
  2      4    -0.1291  -0.1385 -0.1198  True
  2      5     0.0239   0.018   0.0297  True
  2      6    -0.0563  -0.0612 -0.0513  True
  2      7    -0.0822  -0.0875 -0.0769  True
  2      8    -0.1533  -0.1578 -0.1487  True
  3      4    -0.0981  -0.1112  -0.085  True
  3      5     0.0549   0.044   0.0658  True
  3      6    -0.0253  -0.0357 -0.0148  True
  3      7    -0.0512  -0.0619 -0.0406  True
  3      8    -0.1222  -0.1325  -0.112  True
  4      5     0.153    0.1435  0.1625  True
  4      6     0.0729   0.0639  0.0818  True
  4      7     0.0469   0.0377  0.0561  True
  4      8    -0.0241  -0.0329 -0.0154  True
  5      6    -0.0801  -0.0854 -0.0749  True
  5      7    -0.1061  -0.1117 -0.1005  True
  5      8    -0.1771   -0.182 -0.1722  True
  6      7     -0.026  -0.0306 -0.0213  True
  6      8     -0.097  -0.1008 -0.0932  True
  7      8     -0.071  -0.0753 -0.0668  True
---------------------------------------------
[1 2 3 4 5 6 7 8]
```

In depth Analysis – Data analysis, Machine learning & modelling

The dependent variable or the response variable is a discrete variable consisting of eight categories of insurance applicants. Since the response variable is not continuous, linear regression cannot be used for the analysis. And, since the data is categorical in nature, we will use supervised machine learning classifier algorithms for the analysis.

ML Classifier Algorithms

The following are some of the classifier algorithms considered for the analysis.

- Naïve Bayes Classifier
- K nearest neighbors (KNN) classifier
- Support Vector Machines
- Logistic Regression Classifier
- Decision Classifier
- Random Forest Classifier

Support vector machines (SVM) and logistic regression works well when there are only two classes of response variable to be classified such as weather an customer's insurance application is approved or not. Here, the applications are classified into eight categories. Therefore, Naïve Bayes, KNN, Decision Classifier, and Random Forest algorithms are used for the analysis. The model with the best accuracy scores are further studied to optimize the model.

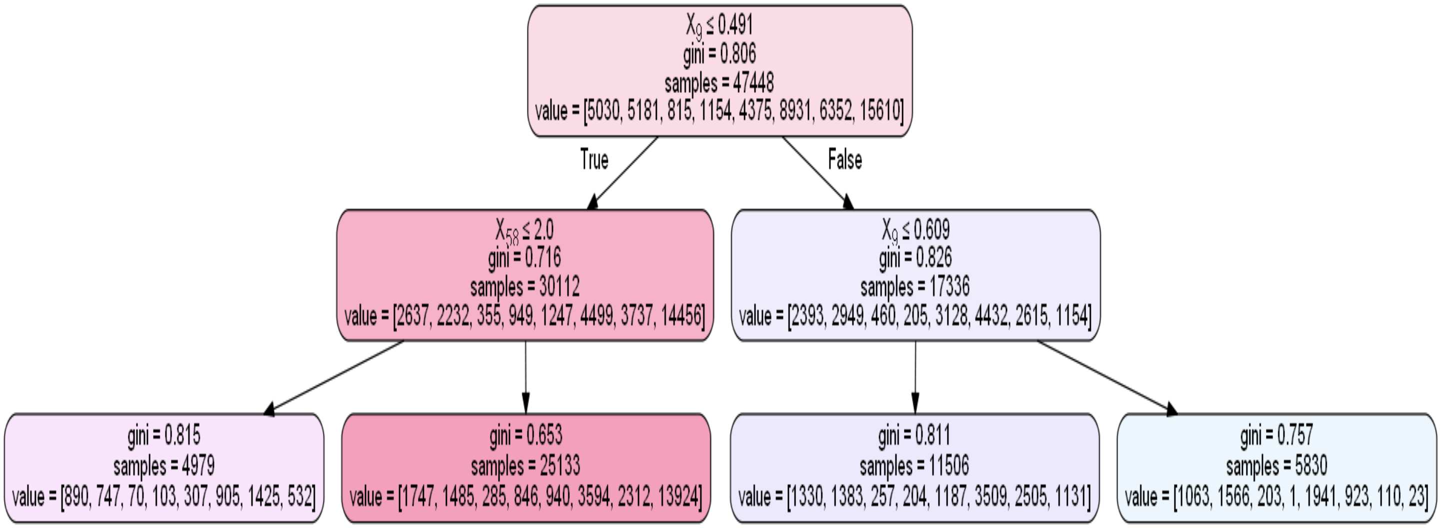
The data is split into test and training set (80/20 split) using the test-train split function and the model is trained with the training data with the classifier default settings. The accuracy scores are obtained as seen in the following table:

Classifier	Accuracy Score
Naïve-Bayes Classifier	0.39
K nearest neighbor(KNN)	0.24
Decision Tree model	0.53
Random Forest Classifier	0.51

From the above table Decision Tree model and the Random Forest Classifier has the best accuracy scores compare to KNN and Naïve Bayes Classifiers. Next, we will explore model optimization and best fit model for the analysis.

Decision Tree visualization

The following diagram is the decision classifier visualization to see how the data is being split at each level. The root node (top most) is the variable BMI, which means it is the most important feature for classification of this dataset. Only two split depths are shown here for easier understanding. In the next section, we see that BMI in fact has the highest influence on the model using the feature importance output of the model.

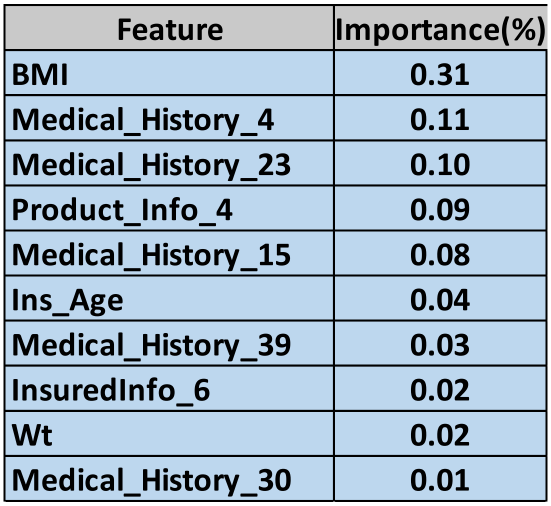

$X_9 \leq 0.491$
gini = 0.806
samples = 47448
value = [5030, 5181, 815, 1154, 4375, 8931, 6352, 15610]
True
False
$X_{58} \leq 2.0$
gini = 0.716
samples = 30112
value = [2637, 2232, 355, 949, 1247, 4499, 3737, 14456]
$X_9 \leq 0.609$
gini = 0.826
samples = 17336
value = [2393, 2949, 460, 205, 3128, 4432, 2615, 1154]
gini = 0.815
samples = 4979
value = [890, 747, 70, 103, 307, 905, 1425, 532]
gini = 0.653
samples = 25133
value = [1747, 1485, 285, 846, 940, 3594, 2312, 13924]
gini = 0.811
samples = 11506
value = [1330, 1383, 257, 204, 1187, 3509, 2505, 1131]
gini = 0.757
samples = 5830
value = [1063, 1566, 203, 1, 1941, 923, 110, 23]

Model Optimization

In this section we explore further using the decision tree classifier for optimization using the following methods.

Model run time optimization using feature importance:

The dataset has more than 100 independent variables. Feature importance ranks the independent variables that are most influential to the classifier model. The following table contains the top ten features (variables) influencing the model.



Feature	Importance(%)
BMI	0.31
Medical_History_4	0.11
Medical_History_23	0.10
Product_Info_4	0.09
Medical_History_15	0.08
Ins_Age	0.04
Medical_History_39	0.03
InsuredInfo_6	0.02
Wt	0.02
Medical_History_30	0.01

From the above table, BMI has the highest influence on the model with 31%. The model is re-run using only the above 10 independent variables instead of the 100 variables. The accuracy scores match the previous model with the same score of 0.53. The advantage is the reduced processing time of the model for the same result. Previously, the model running time was 0.625 and after reducing the independent variables it is only 0.109, a reduction of 5.7X.

Reducing model overfitting using cross validation:

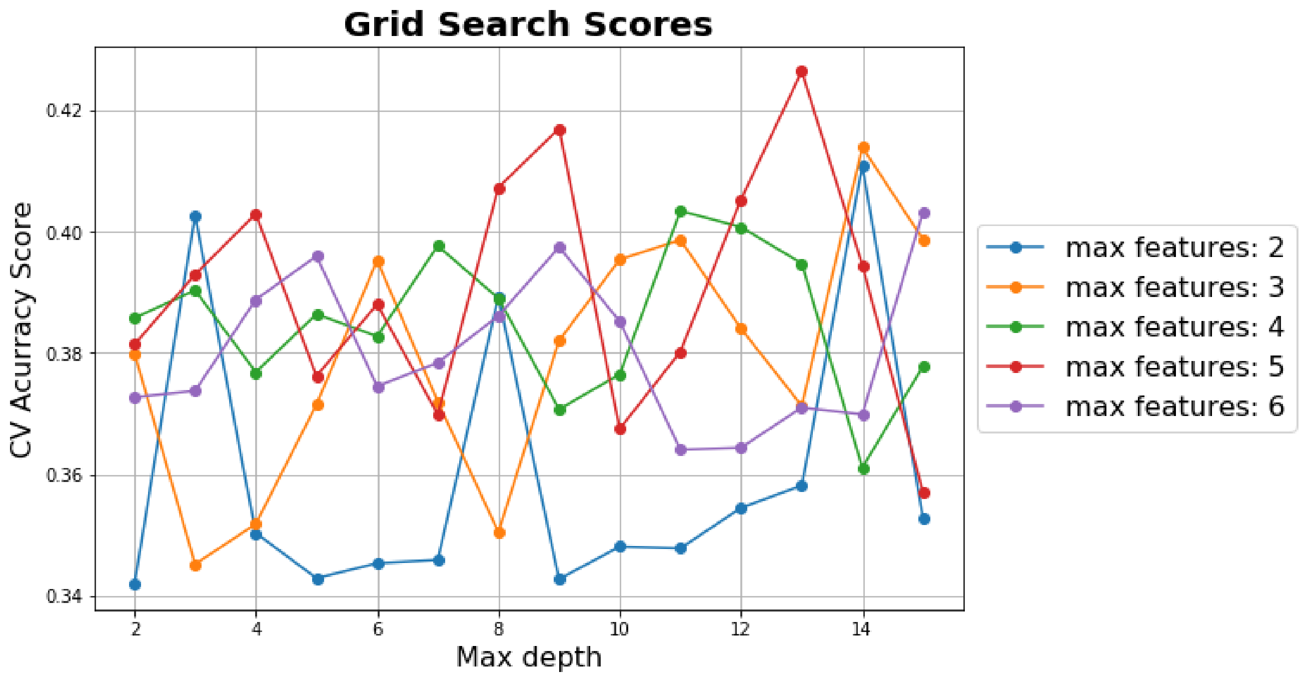
Using cross validation, the training dataset is split into five chunks and cross validated to remove the overfitting of the data. Again, using the original default parameters the accuracy score drops to around 0.41 from 0.53. This is a clear indication of the model overfitting and we would have seen higher error rate in the prediction of the test data if we had not cross validated the training data. The drop in the accuracy score could be due to the quality of the data. Still, 0.41 accuracy score is better than Naïve Bayes or KNN models.

Hyper parameter tuning using Randomized search CV and Grid search CV:

Using Randomized search CV, we get model hyper parameters that gives the best accuracy are determined. The following table has hyper tuned parameters:

Parameters	Values
Split Criterion (gini /entropy)	entropy
Decision Tree Max Depth	14
Max features	8
Minimum Sample leaf	6
Best Score	0.45

Using Grid search CV method, the hyper parameters are obtained plotted as shown in the grid below.



From both the above method we see an optimized accuracy score of around 0.45 without overfitting the model.

Conclusion

From the above study:

- We can conclude that Decision Tree classifier is the best solution for this dataset based on the raw accuracy scores with default parameters on KNN, Naïve Bayes, Decision Tree, and Random Forest models.
- Using feature importance output of the Decision Tree Classifier, the top 10 independent variables affecting the model are chosen and the model is refit with only the top 10 influencer variables resulting with the same accuracy scores.
- Model optimization is done to minimize overfitting using cross validation. Model hyper tuning is done to improve the accuracy scores using randomized search and grid search cross validation.
- The best score for this model is 0.45.