Prudential life insurance data set

Gautham Gowda

Contents

- Problem statement- why is it useful to answer the question
- Clients and intended audience
- Dataset used for the investigation
- Data cleaning and wrangling
- Data visualization
- Exploratory data analysis (EDA)
- Machine learning algorithms
- Conclusions

Motivation for this study

- 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.
- Predicted response is the classification of applicants into eight classes

Clients/ Intended audience

- Any companies that offer insurance or loans to customers based on customer profiles to predict risk.
- It can be applied to a variety of predictive analytics cases

- 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.

Dataset – prudential life data

- The dataset used is Prudential Life Insurance application classified into eight different categories
- https://www.kaggle.com/c/prudential-life-insurance-assessment/overview.

- Data is coded to hide personal information

Dataset- variables list

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

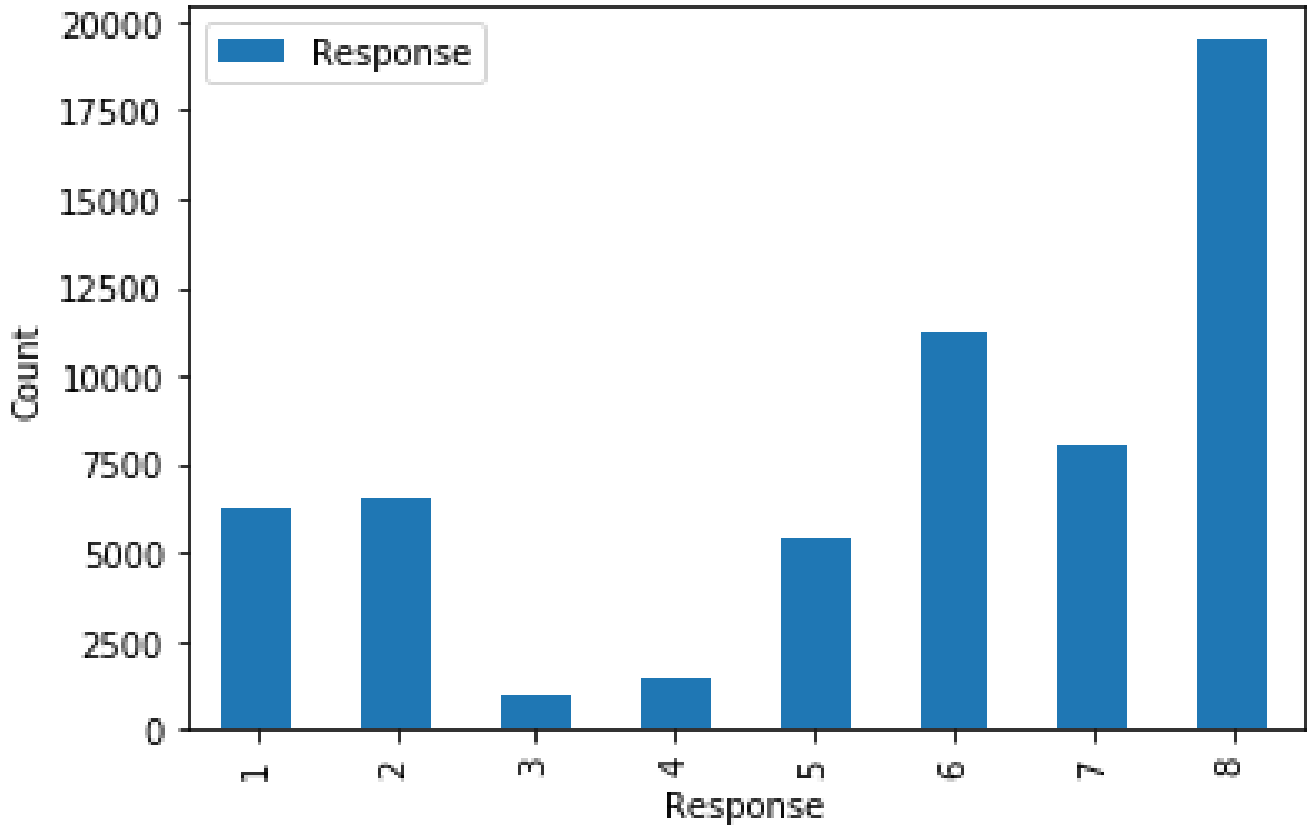
Data cleaning and wrangling

- The variables are arranged as columns
- All the continuous variables are normalized between 0 and 1
- Missing values are filled with the following method:

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

- Outliers in the data is kept since the impact on EDA was minimal.

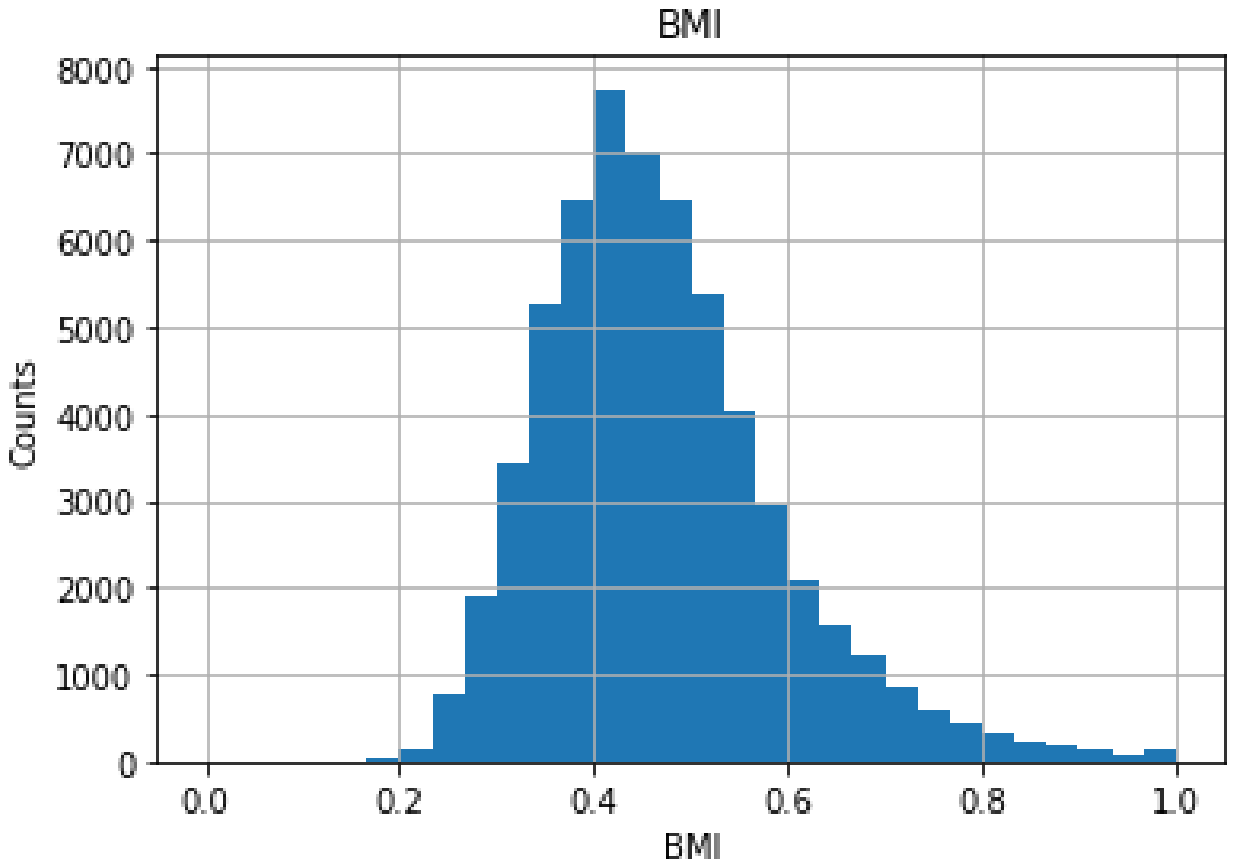
Data visualization – Response



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

Response (dependent variable) is classified into 8 applicant categories

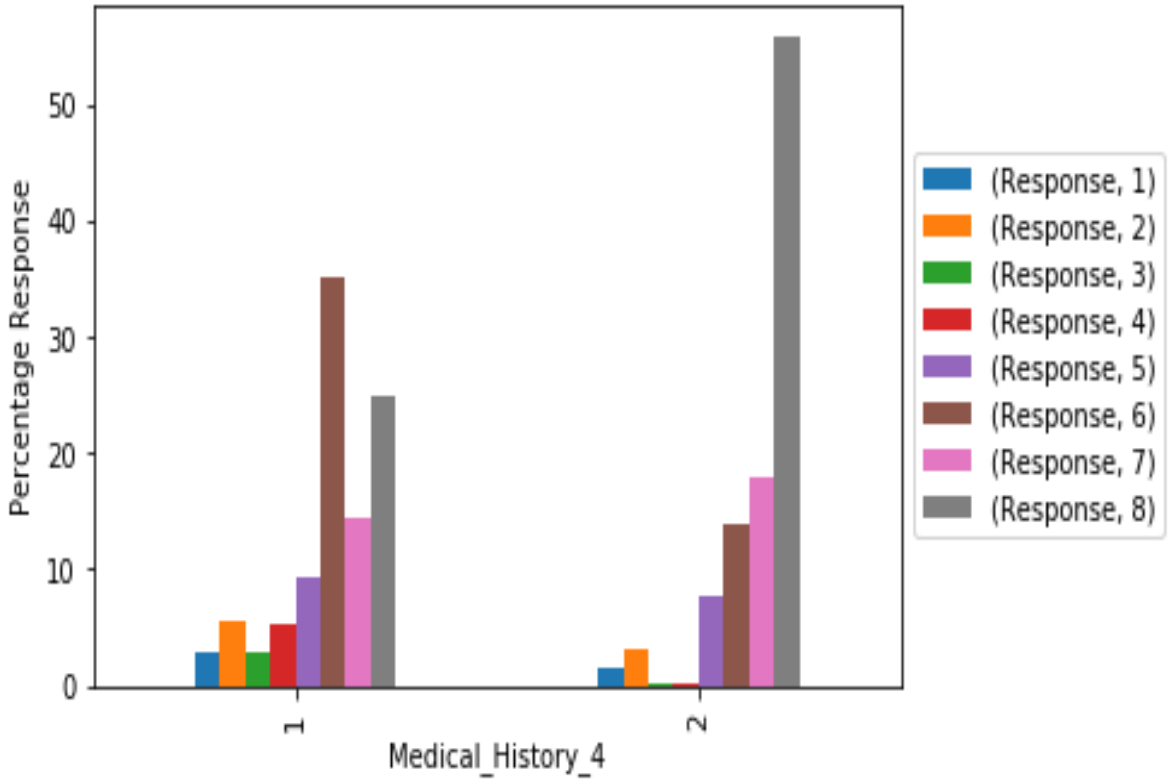
Data visualization (Histogram) – BMI



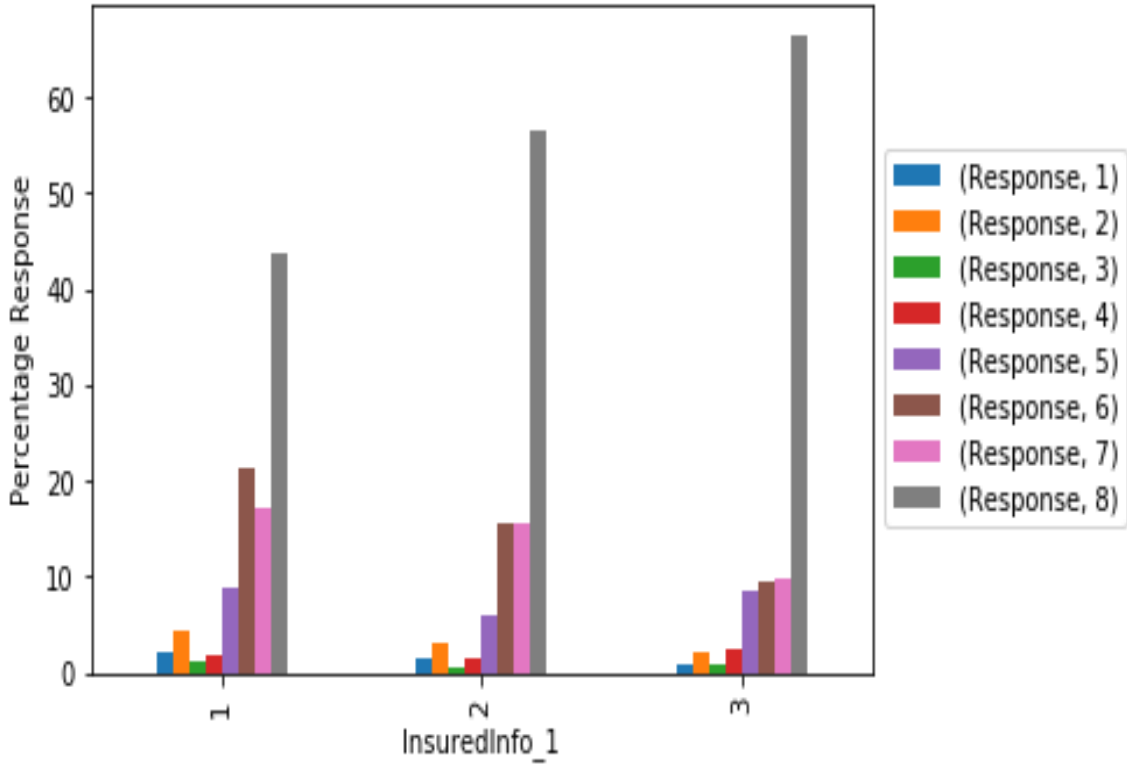
Summary Statistics	BMI
count	59381.000000
mean	0.469462
std	0.122213
min	0.000000
25%	0.385517
50%	0.451349
75%	0.532858
max	1.000000

Continuous variable example -Distribution and summary statistics for BMI

Data visualization – categorical variables



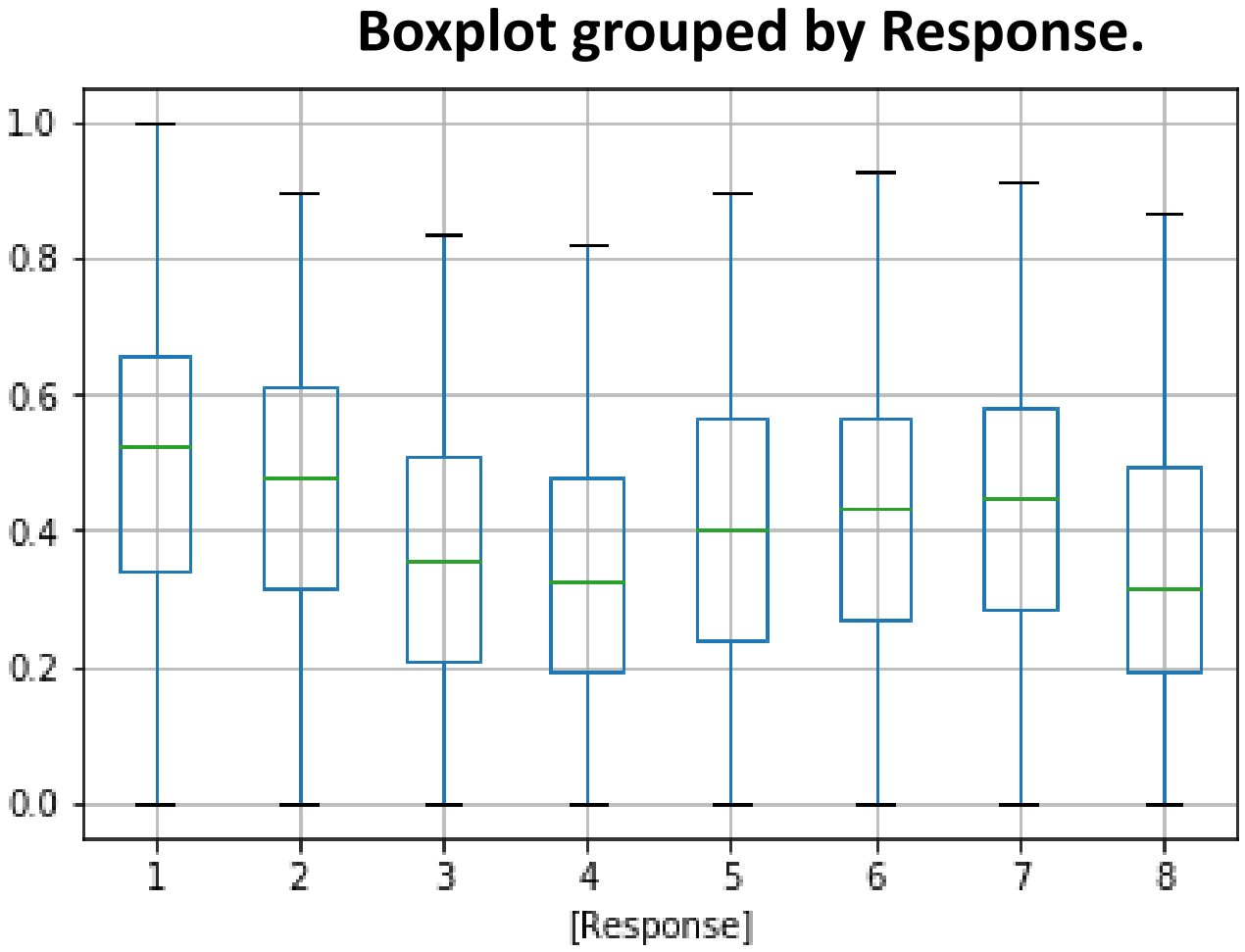
Response v. Medical_History_4



Response v. Insuredinfo_1

Response categories as a percentage of categorical variables

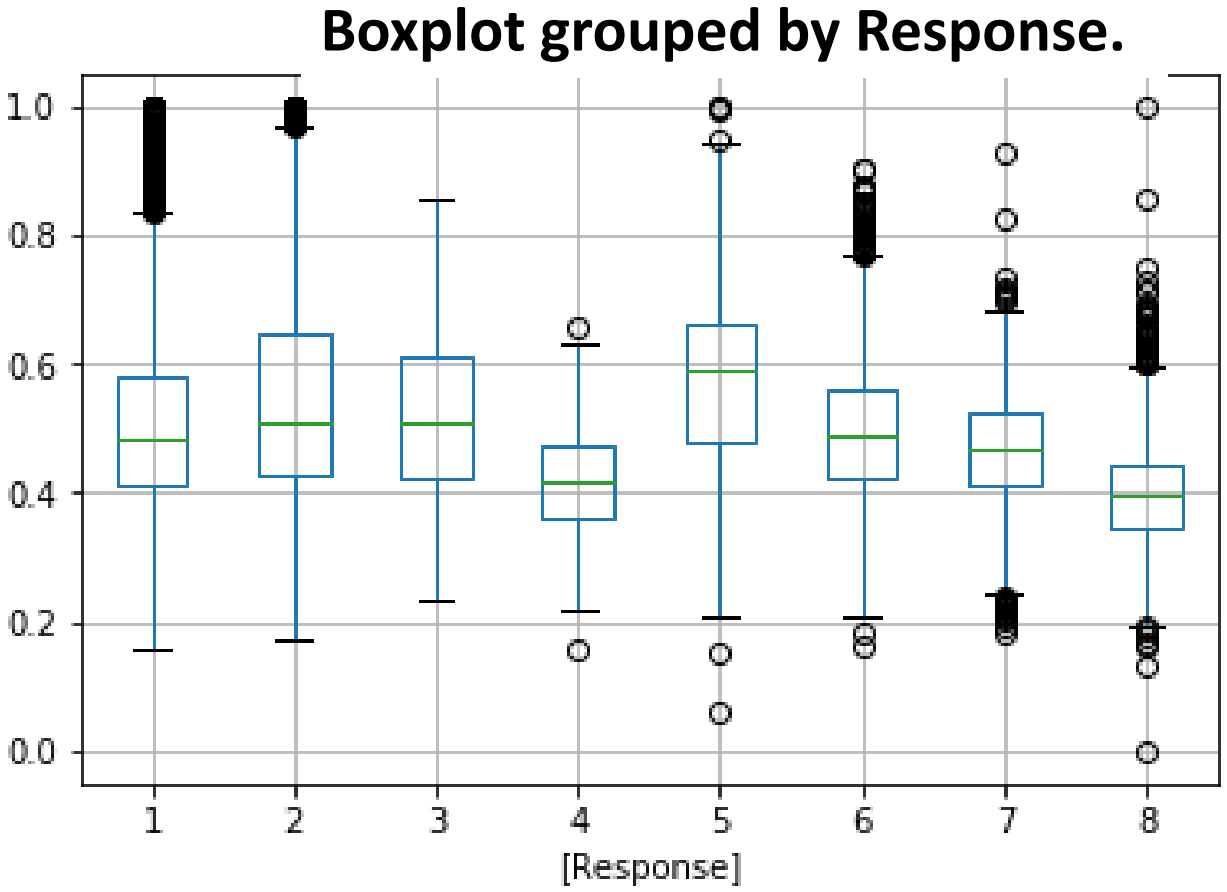
EDA – Age v. Response



Response	mean	std
1	0.492908	0.202704
2	0.460992	0.193211
3	0.359476	0.188468
4	0.337033	0.173476
5	0.404595	0.201250
6	0.426265	0.188431
7	0.434459	0.185575
8	0.342974	0.185573

Box plot shows Response Catogory-4 & 8 has the lowest Age. Category-1 has the highest

EDA – BMI v. Response



Response	mean	std
1	0.509306	0.147606
2	0.546909	0.157569
3	0.515884	0.119768
4	0.417773	0.075936
5	0.570763	0.127321
6	0.490632	0.098622
7	0.464679	0.076837
8	0.393644	0.069260

Box plot shows Response Catogory-8 has the lowest BMI & Category-5 has the highest

EDA – correlation matrix

	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

BMI and Wt. has strong correlation
BMI, Age, Wt. are negatively correlated to Response

EDA – hypothesis tests

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]

- Null Hypothesis (H0): There is no difference in Mean (BMI/Age) between the response categories
- Alternate Hypothesis (H1): There is a difference in Means between responses

Multiple comparison of means at alpha = 0.05 significance suggests data is significant. Reject the null hypothesis

Analysis- machine learning implementation

- Insurance applicant responses are classified into eight responses – categorical analysis

- Evaluate different ML classifiers and choose the algorithm with best accuracy score as a recommendation

- ML Algorithms considered
  - KNN nearest neighbors
  - Naïve Bayes
  - Support Vector Machines (SVM)
  - Logistics Regression
  - Decision Classifier
  - Random Forest

Machine learning – models comparison

- Various classifier algorithms used and accuracy scores are obtained
- SVM and logistic regression classifiers will not work for our data since there are more than 2 response classes

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

Decision tree & Random forest classifier has good accuracy scores

Machine learning – Decision Classifier

Decision tree with max depth =2 (for visualization of split)

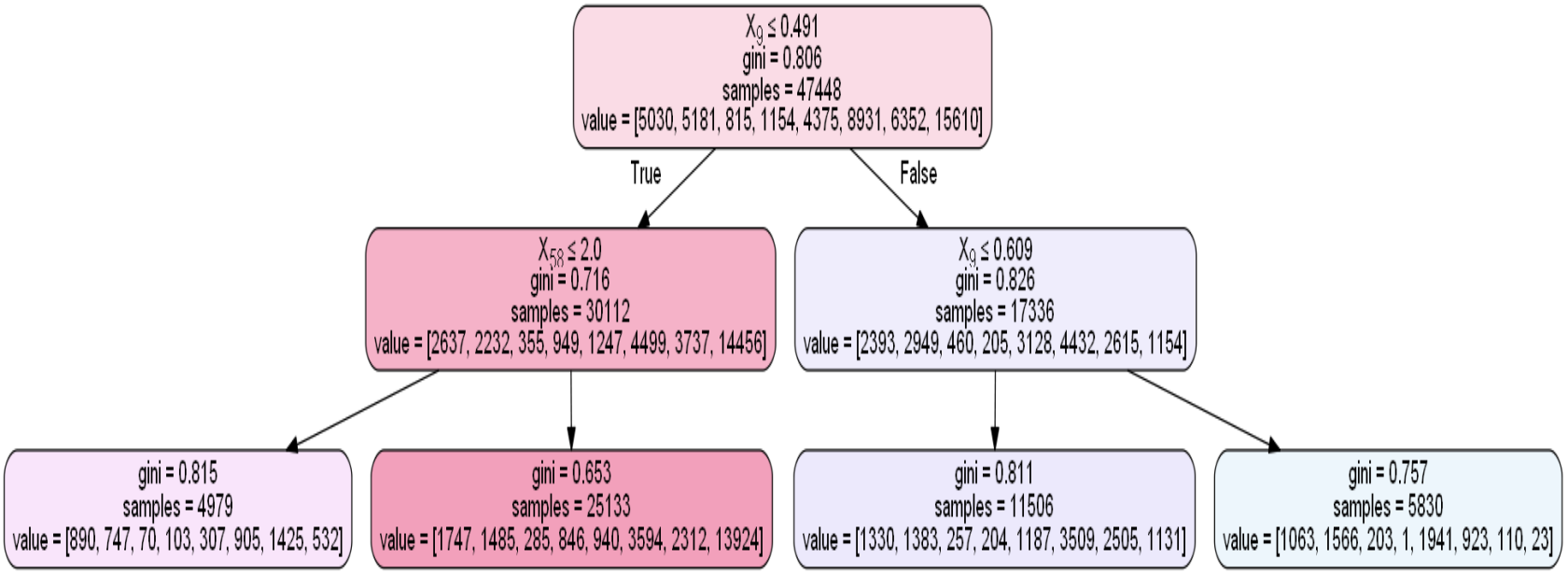


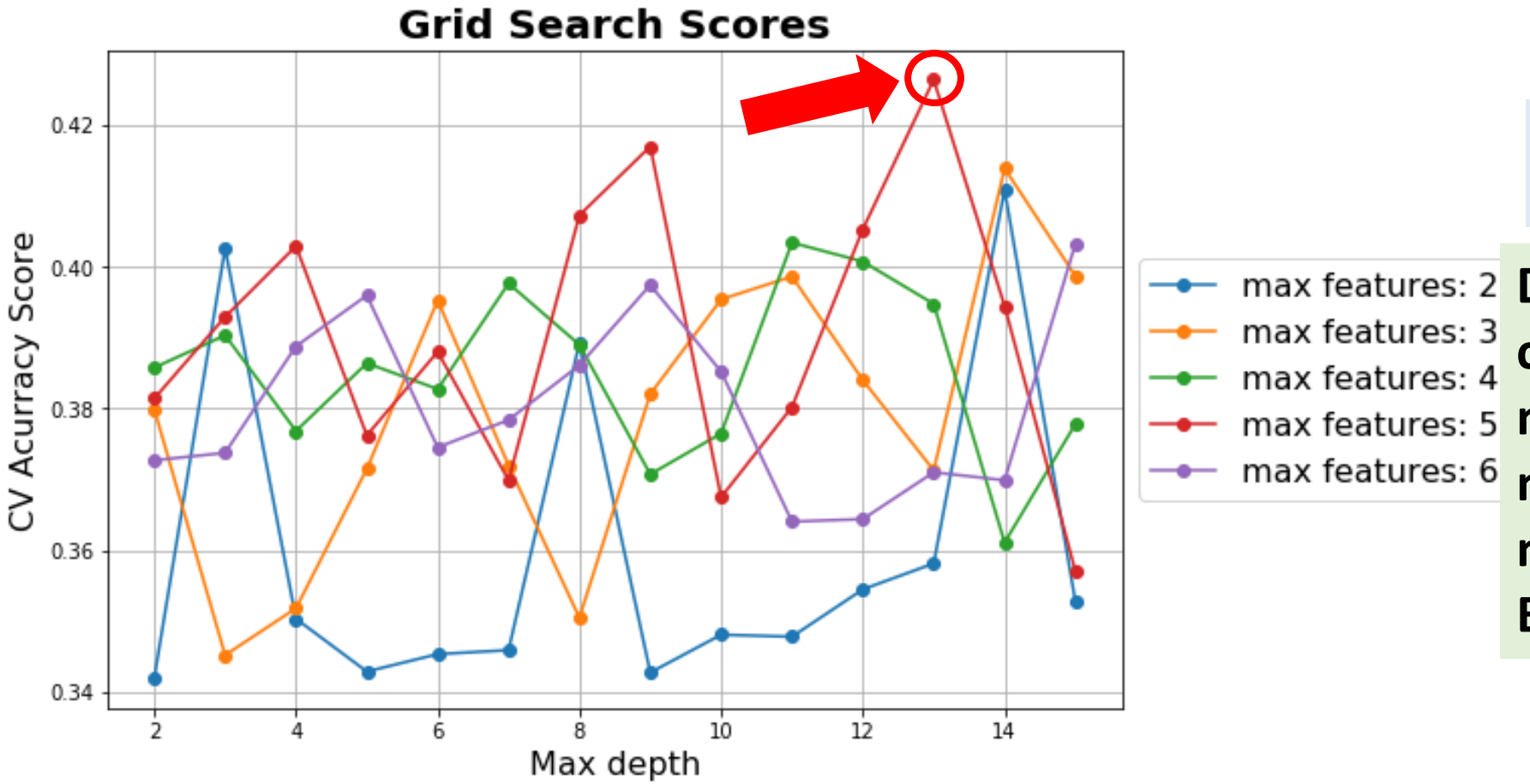
Table: Top 10 variables and their effect on the decision tree classifier

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

Root node is BMI as seen from tree diagram

BMI has 31% effect on decision classifier as seen from the table

Classifier – Cross Validation Accuracy



Hyper-parameter tuning using RandomizedSearchCV

Decision Tree Parameters:
criterion: entropy
max_depth: 13
max_features: 8
min_samples_leaf: 6
Best score is .45

Cross Validation Accuracy at max depth= 13 and max features= 5

Decision classifier model optimization

- Get Feature importance using DecisionTreeClassifier()
  - Get the top 10 independent variables influencing the decision tree
  - Re-run the model with only the 10 features improves the processing time from 0.625 to 0.1094 → improvement of 5.7X

- Cross- Validation of training set to minimize overfitting
  - 5 split cross validation of training set

- Hyper parameter tuning RandomizedSearchCV()
  - criterion: gini, max_depth: 13, max_features: 8 min_samples_leaf: 6
  - run grid search scores to validate the tuned hyper parameters

Conclusion

- The data set is a classification problem
- Classifier algorithms considered: Naïve Bayes, KNN, SVM, Logistic Regression, Decision Tree, Random Forest
- Based on the accuracy scores obtained, Decision Tree Classifier model is chosen to train the data
- BMI is the root node of the decision tree
- Cross validation of the training data is done with CV module
- Hyper parameters tuning with max_depth =14 and max_features = 6 is chosen