LIFE INSURANCE

What is Life Insurance?

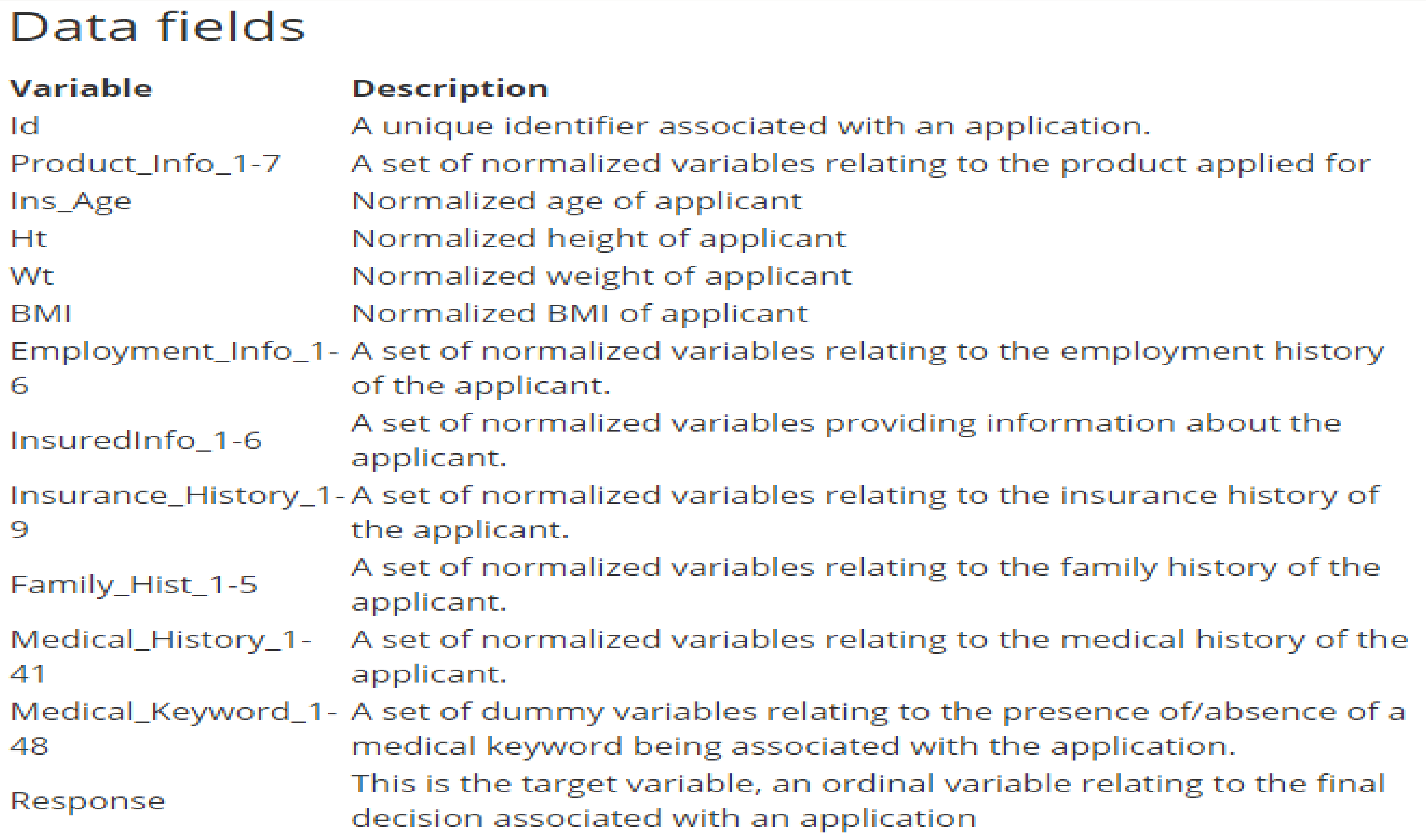
- Life Insurance can be termed as an agreement between the policy owner and the insurer, where the insurer for a consideration agrees to pay a sum of money upon the occurrence of the insured individual's or individuals' death or other event, such as terminal illness, critical illness or maturity of the policy.

Why to have a Life Insurance?

- Protection
- Liquidity
- Tax Relief
- Money when you need it

CAN YOU MAKE BUYING LIFE INSURANCE EASIER ?

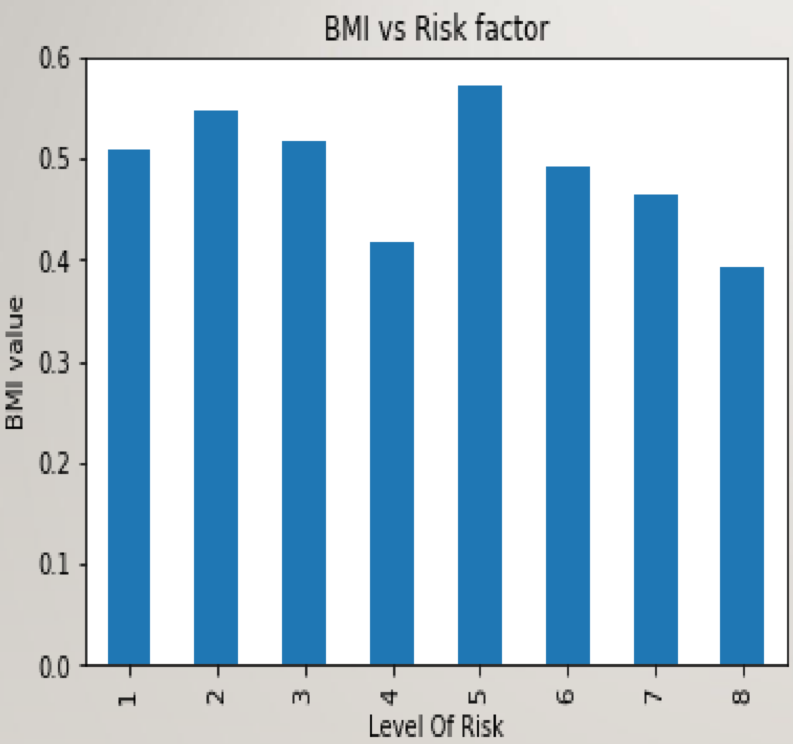
Dataset information: The dataset consists of roughly 59300 customers records and their 126 respective features and one target variable which describe the level of risk.

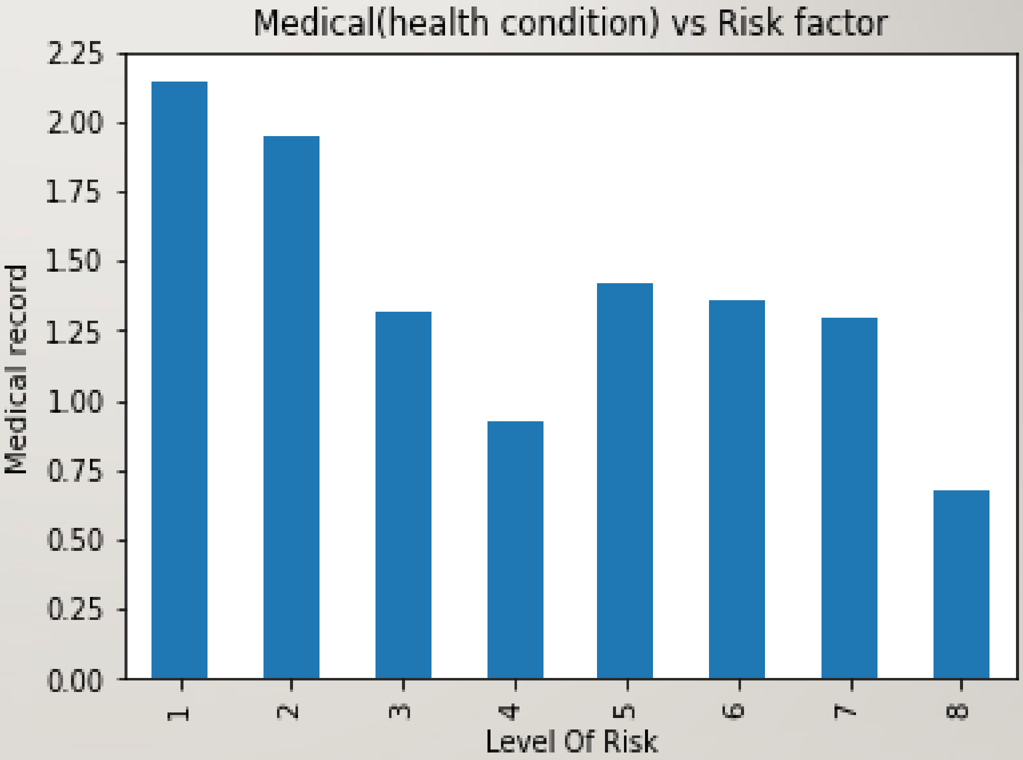


Data fields

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

Dataset Insight:



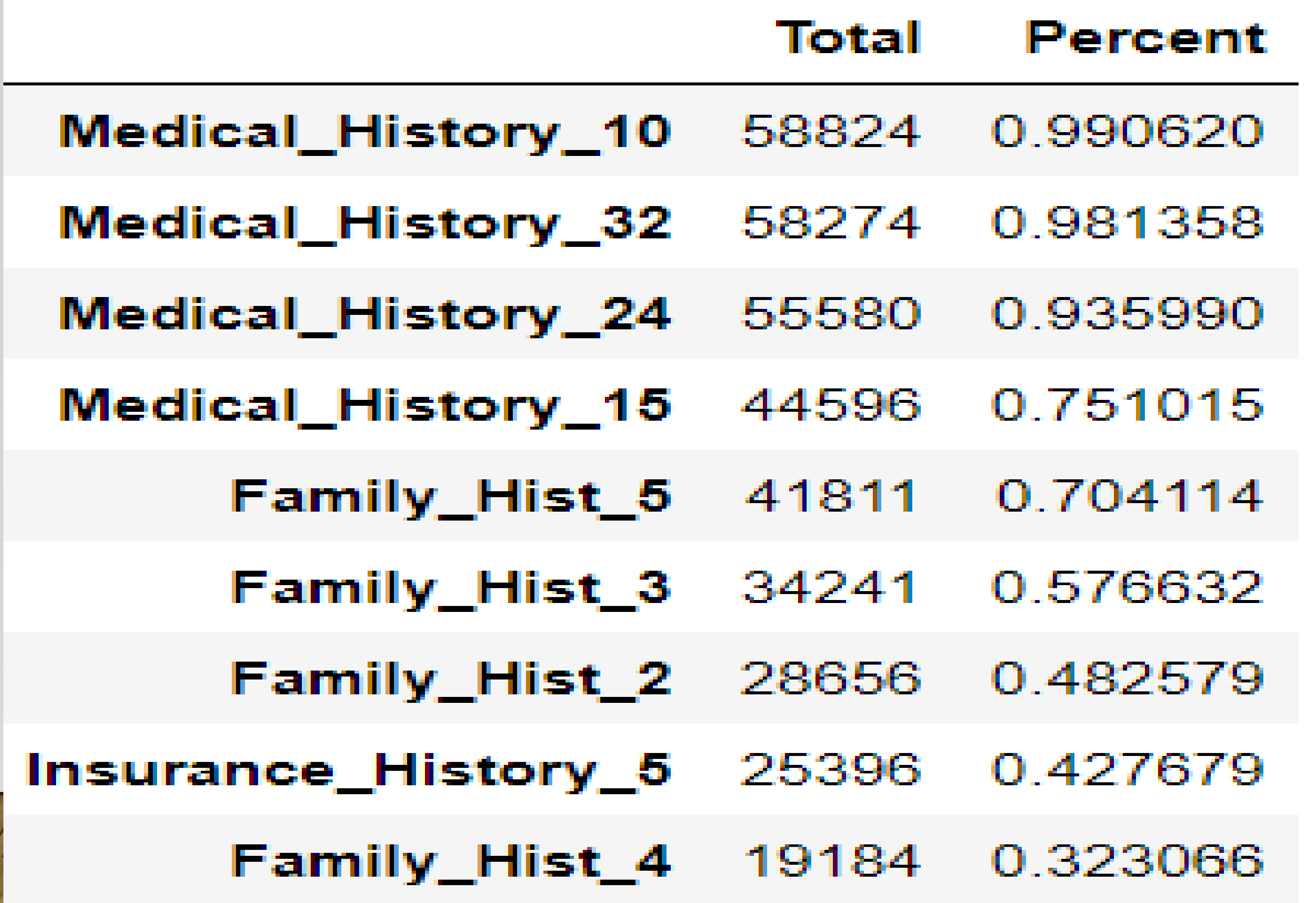


As BMI values are normalized so we can’t take this value to decide the level of risk to company

Higher the medical record higher the risk to insurance company

Feature Engineering:

- Removing columns (Medical_History_10, Medical_History_24 and Medical_History_32) having missing values more than 90%.



	Total	Percent
Medical_History_10	58824	0.990620
Medical_History_32	58274	0.981358
Medical_History_24	55580	0.935990
Medical_History_15	44596	0.751015
Family_Hist_5	41811	0.704114
Family_Hist_3	34241	0.576632
Family_Hist_2	28656	0.482579
Insurance_History_5	25396	0.427679
Family_Hist_4	19184	0.323066

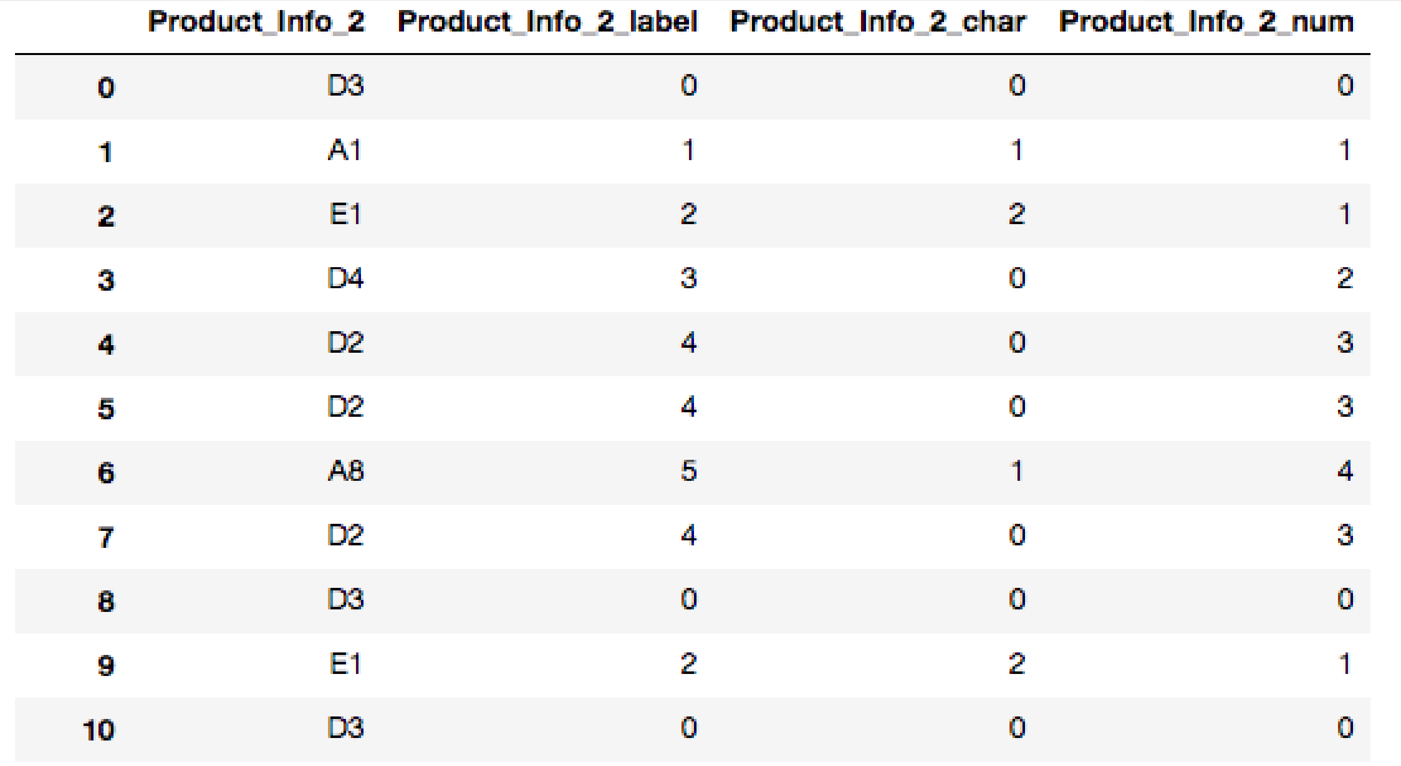
Feature Engineering:

- There is a categorical variable called Product_Info_2 which contains character and number. I had factorize the column and split the character and number, then create additional two columns with the extract character and number after factorization.

Label encoded of column

Label encoded of only character

Label encoded of only number

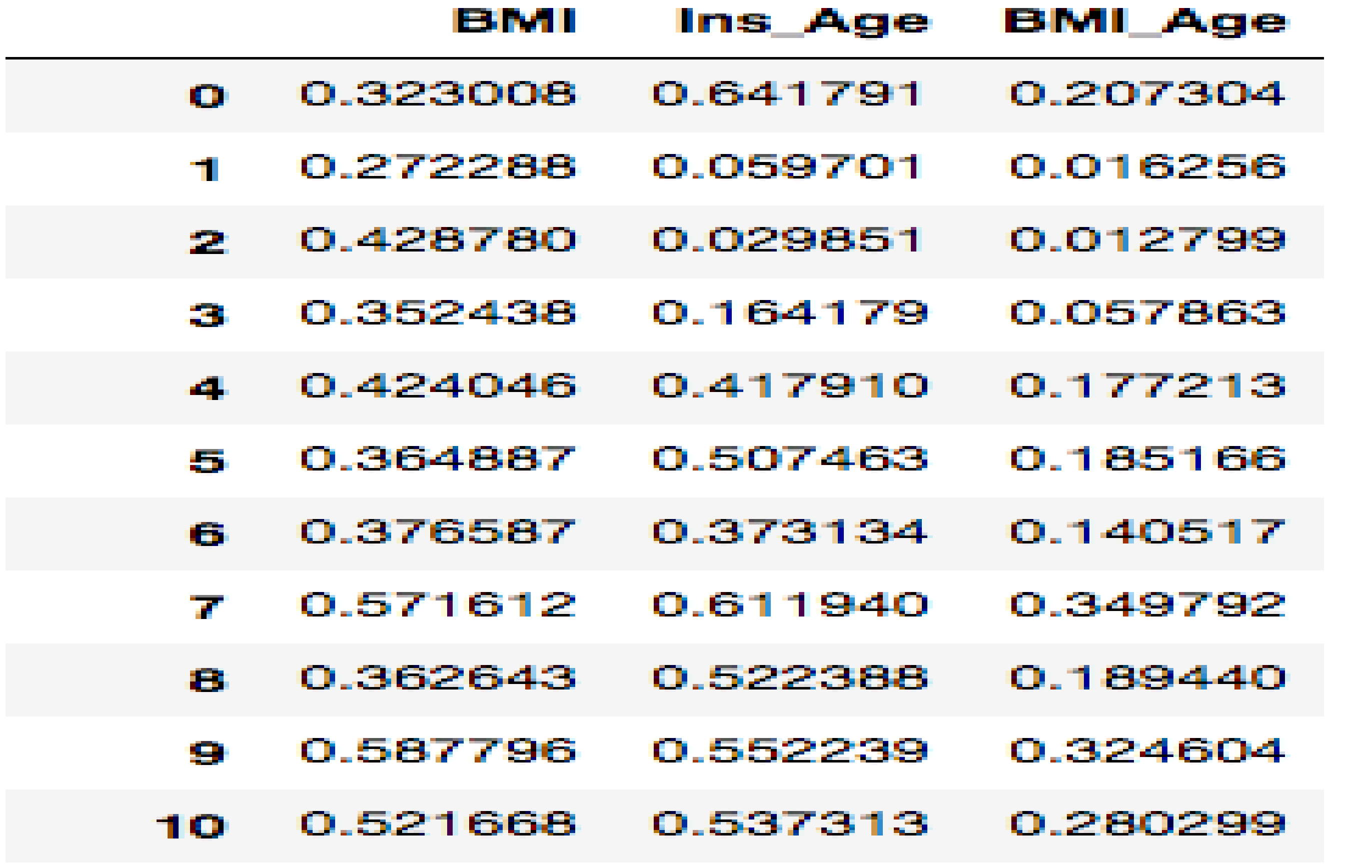


	Product_Info_2	Product_Info_2_label	Product_Info_2_char	Product_Info_2_num
0	D3	0	0	0
1	A1	1	1	1
2	E1	2	2	1
3	D4	3	0	2
4	D2	4	0	3
5	D2	4	0	3
6	A8	5	1	4
7	D2	4	0	3
8	D3	0	0	0
9	E1	2	2	1
10	D3	0	0	0

Feature Engineering:

- Created a new features by multiply the BMI column and Ins_Age column value because the product these two feature having same significant since & it is a useful feature for model to learn.

Product of BMI & Ins_Age columns

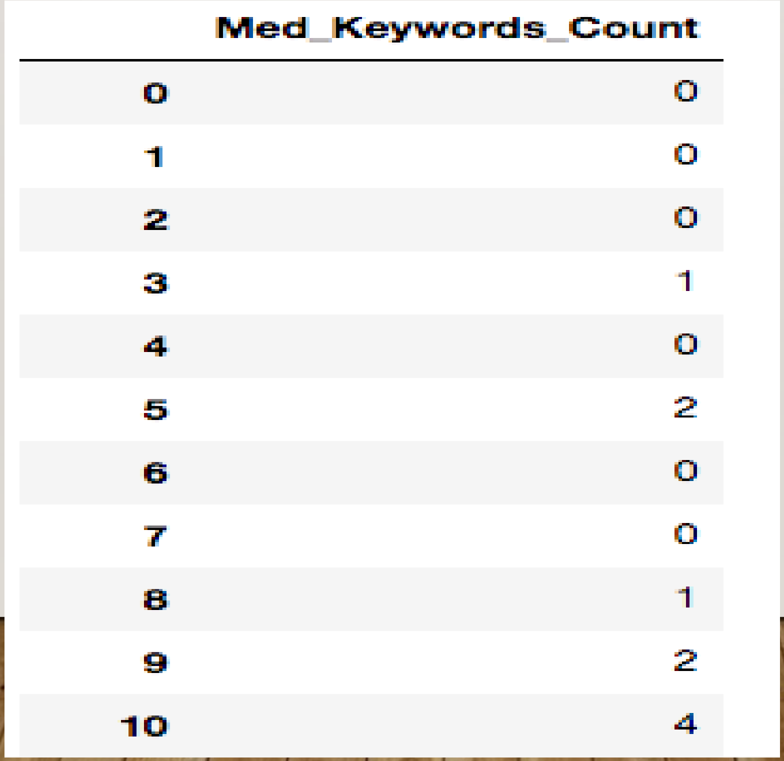


	BMI	Ins_Age	BMI_Age
0	0.323008	0.641791	0.207304
1	0.272288	0.059701	0.016256
2	0.428780	0.029851	0.012799
3	0.352438	0.164179	0.057863
4	0.424046	0.417910	0.177213
5	0.364887	0.507463	0.185166
6	0.376587	0.373134	0.140517
7	0.571612	0.611940	0.349792
8	0.362643	0.522388	0.189440
9	0.587796	0.552239	0.324604
10	0.521668	0.537313	0.280299

Feature Engineering:

- For the Medical_Keyword columns, it has 48 in totals and it is a set of dummy variables relating to the presence of/absence of a medical keyword being associated with the application. I added a column which sum all the counts of those dummy variables.

Sum of all 48 Medical_counts



	Med_Keywords_Count
0	0
1	0
2	0
3	1
4	0
5	2
6	0
7	0
8	1
9	2
10	4

Machine Learning Model:

No.	Model	quadratic_weighted_kappa
1	**Logistic Regression**	**0.3092**
2	**XGBClassifier**	**0.6239**
3	**RandomForestClassifier**	**0.6995**

Random Forest model is giving the best performance among the models considered and is used for the final prediction.

**XGBClassifier with GridSearchCV is overfitting the data.

GET INSURED

Thank You

SADIQ
Data Scientist