

Life Insurance Risk Prediction using Machine Learning Algorithms`

AGENDA



Industry research



Data problem and description




Data Understanding [EDA]



Model



Data interpretation



Business application

Industry research

- Insurance companies growth reported: 8.6% growth in Life insurance and Annuities in the US market.
- Thus, reducing the **challenges** associated with life insurance application and **saving costs** through greater automation while making sure that the accuracy of risk assessment is not compromised.
- Opportunities to solve complex challenges
 - processing time involved in the application
 - identifying opportunities that can improve the overall experience
- Chosen insurance company: **Prudential Financial**

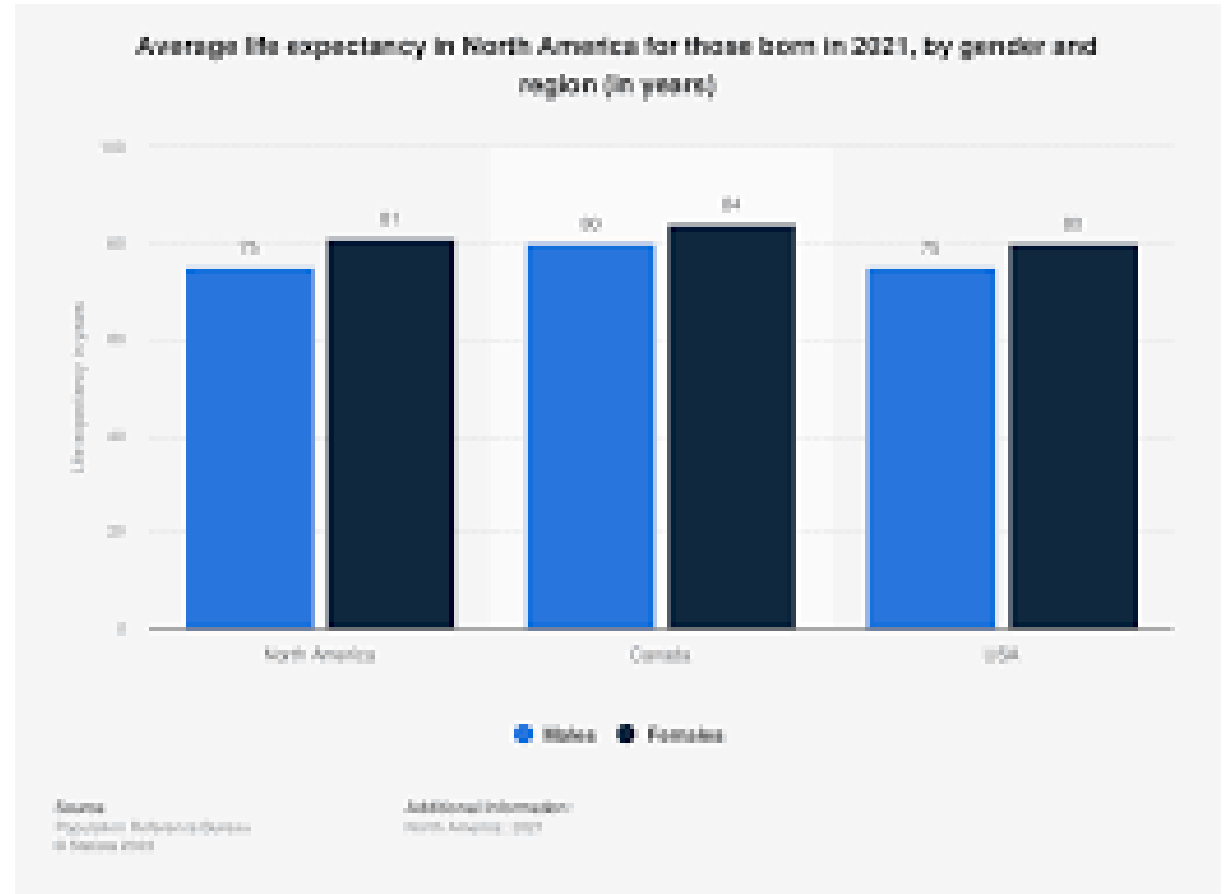


Figure 1. Average life expectancy [2021 US census data]
[Male: 73.2 years; Female: 79.1 years]

Industry research

Prudential Financial

- One of the largest issuers of life insurance in the USA
- Showcase skills in understanding the problem and an attempt in providing a best model that can produce outcomes which can be used to improve performance, yield, risk reduction and enhance business productivity.
- Complex dataset.

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 app
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 applica
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 applican
Medical_Keyword_1-48	A set of dummy variables relating to the presence of/absence of a medical with the application.
Response	This is the target variable, an ordinal variable relating to the final decision a

Problem statement

- **Extensive information** is available for risk prediction via the information from customers regarding the medical history, employment history, applicant information etc.
- The goal of this project to **integrate machine learning models** for **risk management** is to save time and money on repetitive and recurrent tasks.
- Accurately integrated model can **reduce processing time** by 30 days and effectively **increase** customer satisfaction.

EDA

Data shape - (59381 rows, 128 features)

Columns: Id, Product info, Height, Weight, Employment history etc.

Target: Response [1-8]; Multiclass classification.

Data description

Missing values and Data imputation

```
datafile.describe()
```

	Id	Product_Info_1	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	
count	59381.000000	59381.000000	59381.000000	59381.000000	59381.000000	59381.000000	59381.000000	59381.000000	59381.000000
mean	39507.211515	1.026355	24.415655	0.328952	2.006955	2.673599	1.043583	0.405567	
std	22815.883089	0.160191	5.072885	0.282562	0.083107	0.739103	0.291949	0.197190	
min	2.000000	1.000000	1.000000	0.000000	2.000000	1.000000	1.000000	0.000000	
25%	19780.000000	1.000000	26.000000	0.076923	2.000000	3.000000	1.000000	0.238806	
50%	39487.000000	1.000000	26.000000	0.230769	2.000000	3.000000	1.000000	0.402985	
75%	59211.000000	1.000000	26.000000	0.487179	2.000000	3.000000	1.000000	0.567164	
max	79146.000000	2.000000	38.000000	1.000000	3.000000	3.000000	3.000000	1.000000	

	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
Employment_Info_6	10854	0.182786
Medical_History_1	8889	0.149694
Employment_Info_4	6779	0.114161
Employment_Info_1	19	0.000320
Medical_History_8	0	0.000000
Medical_History_9	0	0.000000

EDA

Basic data exploration

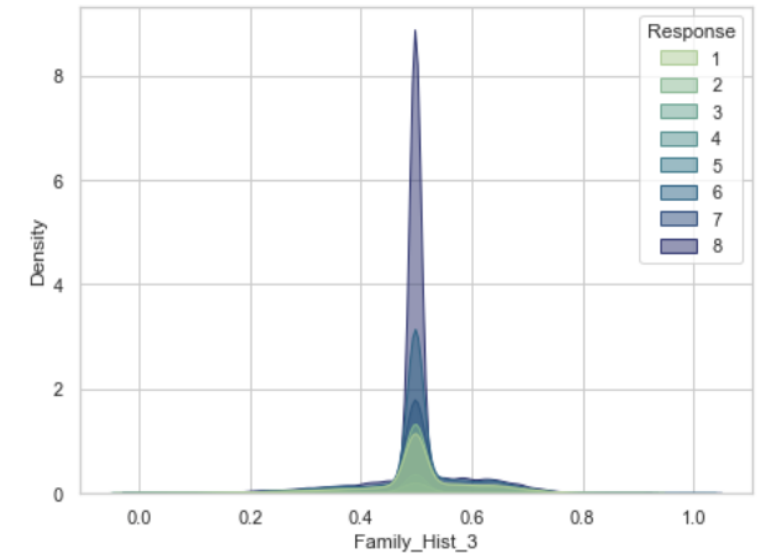
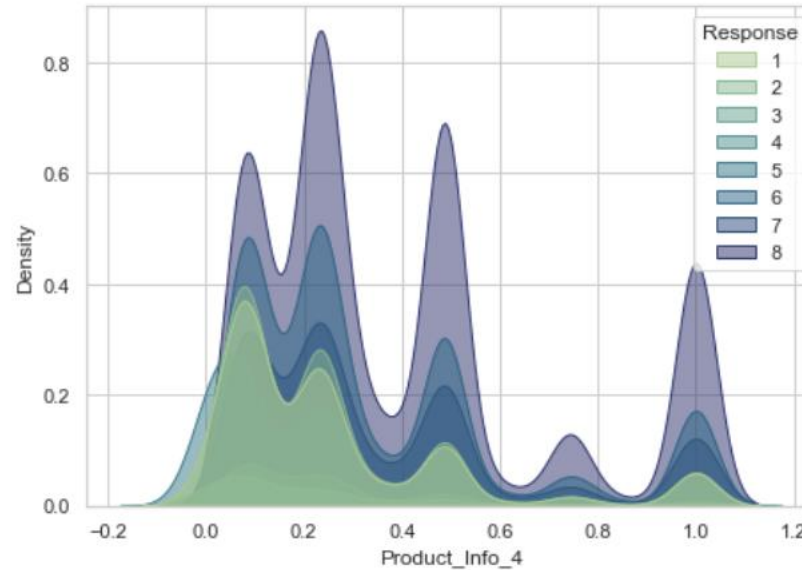
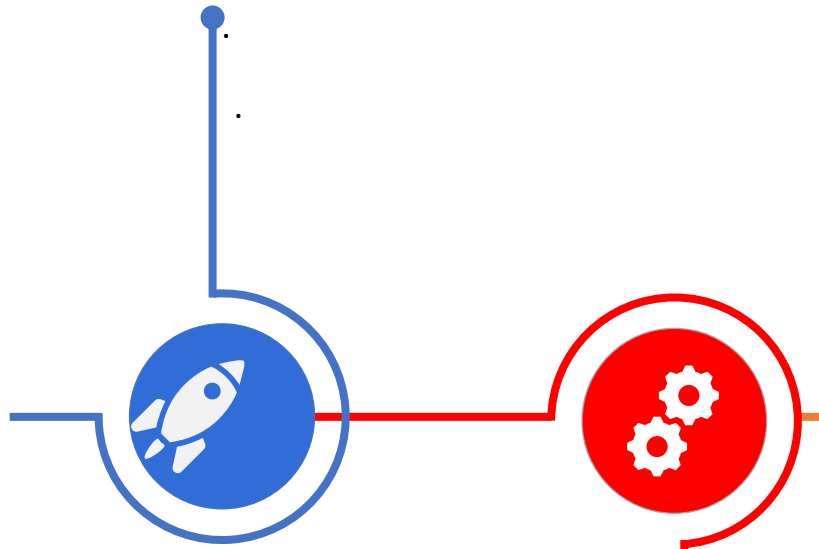


Figure 2: Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function (PDF) of a random variable.

KDE plots

Correlation coefficient

EDA

Spearman rank correlation between the target and features

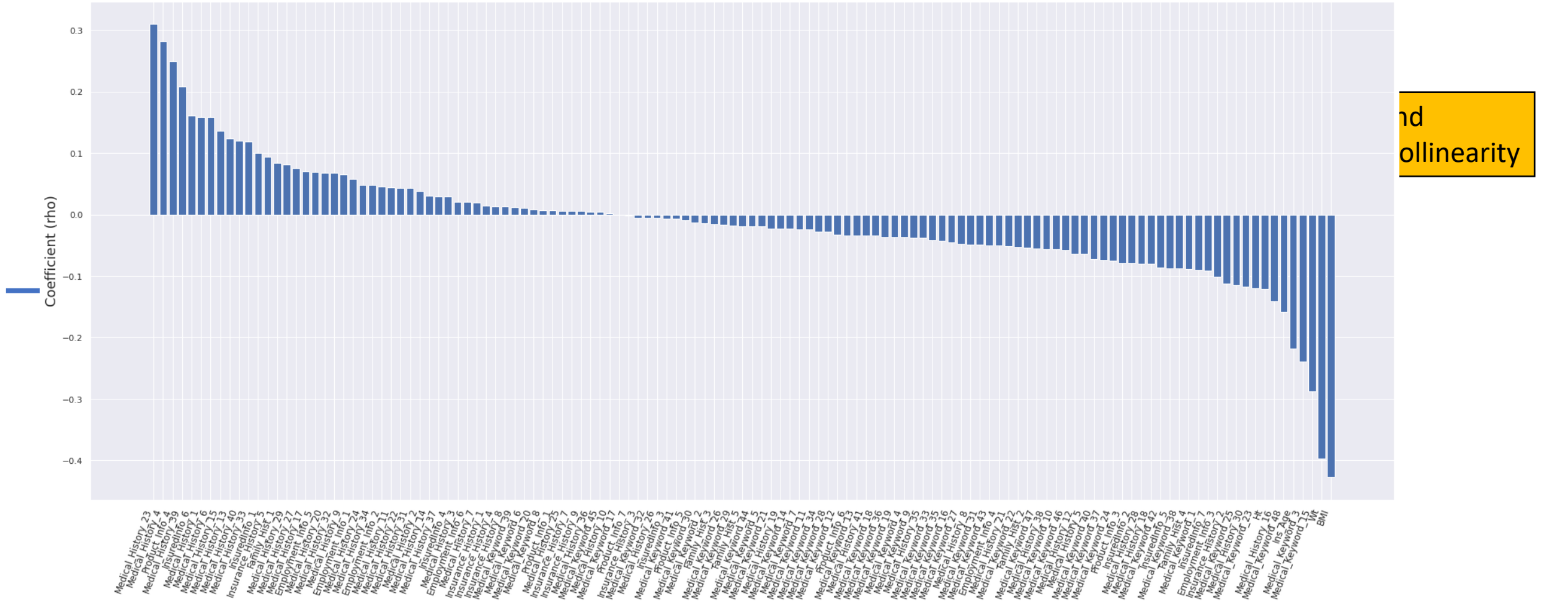
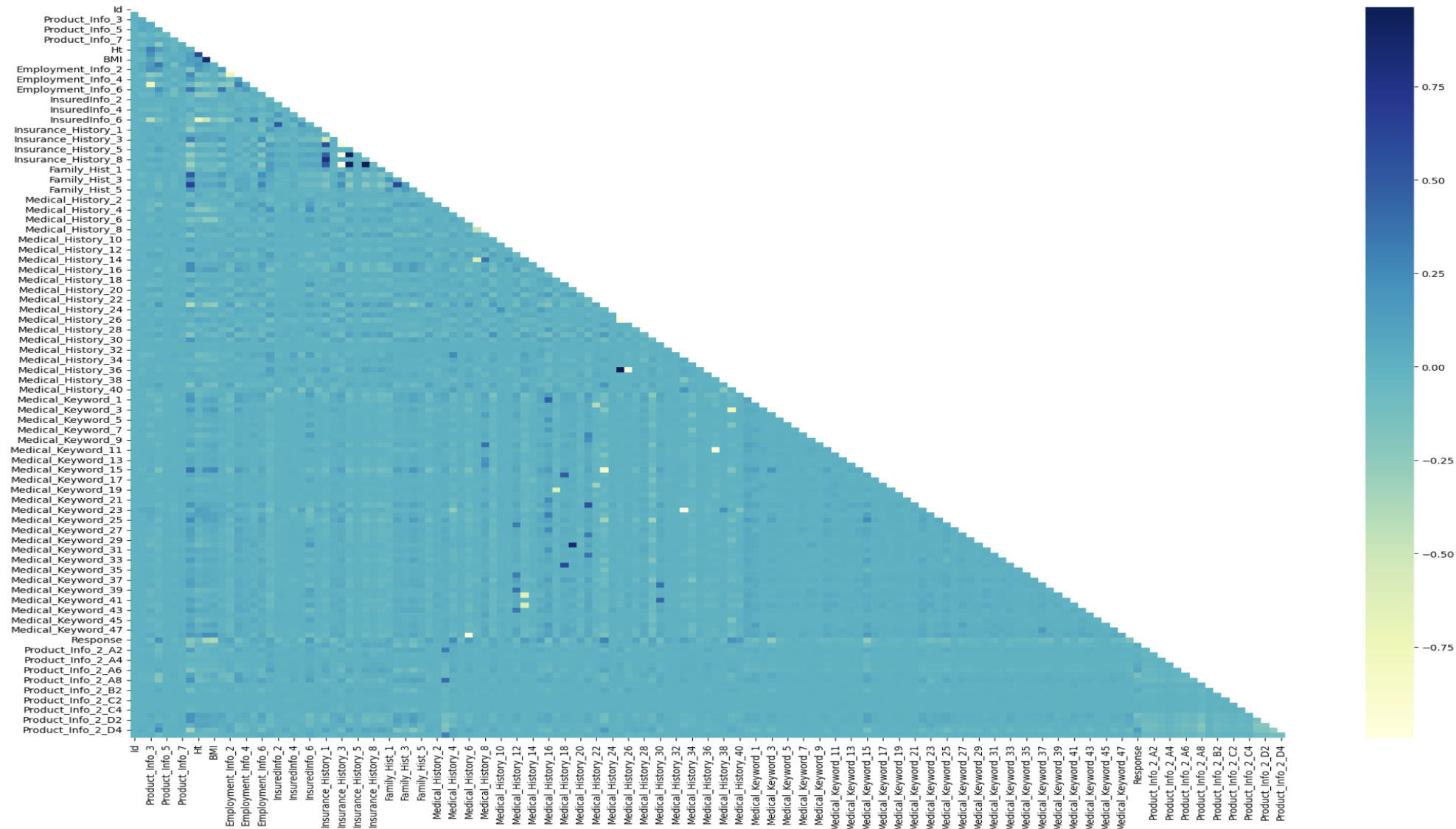


Figure 3: Spearman rank correlation showing relation between features and target variables.


```
#Correlation matrix
correlation_matrix = data.corr()
mask = np.zeros_like(correlation_matrix)
mask[np.triu_indices_from(mask)] = True
plt.figure(figsize=(24,16))
sns.heatmap(correlation_matrix, cmap="YlGnBu", mask = mask)
plt.show()
```



PCA and
multicollinearity



Figure 4: Correlation matrix

Revisit the approach!

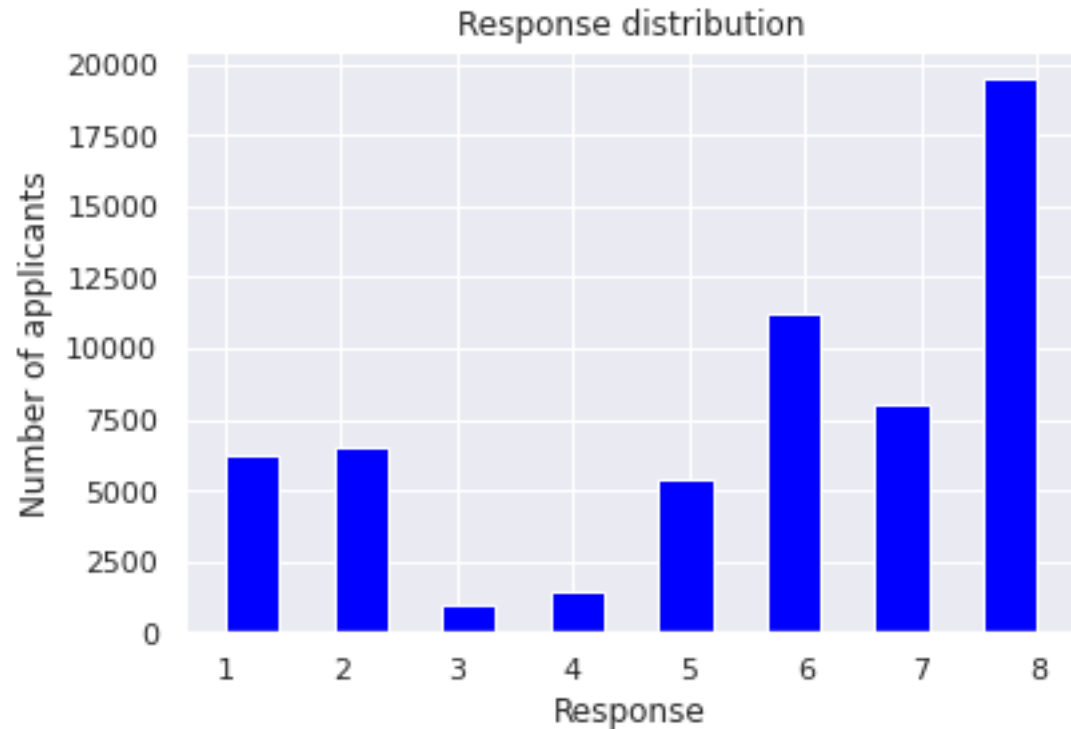


Figure 5: Response distribution

Class balancing!

```
from collections import Counter
from imblearn.over_sampling import SMOTE
oversample = SMOTE()
X_over, y_over = oversample.fit_resample(X_train, y_train)
print(Counter(y_over))
```

Scaling

```
#Scaling of all the features in the dataset
from sklearn.preprocessing import MinMaxScaler

Scaler = MinMaxScaler()
Scaler.fit(X_over)

scale_X_train = Scaler.transform(X_over)
scale_X_test = Scaler.transform(X_test)
```

Model selection

Random forest

Classification Report

	precision	recall	f1-score	support
1	0.35	0.19	0.25	2071
2	0.39	0.22	0.28	2146
3	0.41	0.52	0.46	332
4	0.45	0.69	0.54	451
5	0.55	0.55	0.55	1816
6	0.48	0.46	0.47	3715
7	0.42	0.40	0.41	2580
8	0.68	0.87	0.76	6485
accuracy			0.55	19596
macro avg	0.47	0.49	0.47	19596
weighted avg	0.52	0.55	0.52	19596

Report 1: Classification report for random forest.

Model selection

Extra tree classifier

Classification Report				
	precision	recall	f1-score	support
1	0.39	0.21	0.27	2071
2	0.40	0.24	0.30	2146
3	0.43	0.54	0.48	332
4	0.49	0.64	0.56	451
5	0.56	0.52	0.54	1816
6	0.51	0.51	0.51	3715
7	0.43	0.45	0.44	2580
8	0.69	0.86	0.76	6485
accuracy			0.56	19596
macro avg	0.49	0.50	0.48	19596
weighted avg	0.54	0.56	0.54	19596

Report 2: Classification report for Extra tree classifier.

Model selection

Logistic Regression

Classification Report					
	precision	recall	f1-score	support	
1	0.37	0.28	0.32	2071	
2	0.33	0.20	0.25	2146	
3	0.32	0.46	0.38	332	
4	0.39	0.60	0.47	451	
5	0.45	0.38	0.41	1816	
6	0.43	0.39	0.41	3715	
7	0.41	0.32	0.36	2580	
8	0.64	0.86	0.74	6485	
accuracy			0.51	19596	
macro avg		0.42	0.44	0.42	19596
weighted avg		0.48	0.51	0.48	19596

Report 3: Classification report for Logistic Regression.

Model selection

Decision classifier

Classification Report					
	precision	recall	f1-score	support	
1	0.20	0.22	0.21	2071	
2	0.21	0.23	0.22	2146	
3	0.27	0.42	0.33	332	
4	0.40	0.49	0.44	451	
5	0.36	0.39	0.38	1816	
6	0.38	0.35	0.37	3715	
7	0.28	0.31	0.30	2580	
8	0.69	0.61	0.65	6485	
accuracy			0.41	19596	
macro avg	0.35	0.38	0.36	19596	
weighted avg	0.43	0.41	0.42	19596	

Report 4: Classification report for Decision classifier.

Future studies and Business perspective

Future Studies

- Further hyper tuning to improve F1 score and accuracy*
- Cross validation
- ROC and AUC [scores & curve]
- Lazy predict*

Business Perspective

- Machine Learning can be extensively applied in insurance companies. Examples:
 - Fraud Detection (for non-structured and semi-structured datasets)
 - Customer service (uses supervised and unsupervised machine learning algorithms),
 - Price optimization (the Generalized linear model)
 - Risk Management and other applications

Thank you!!!