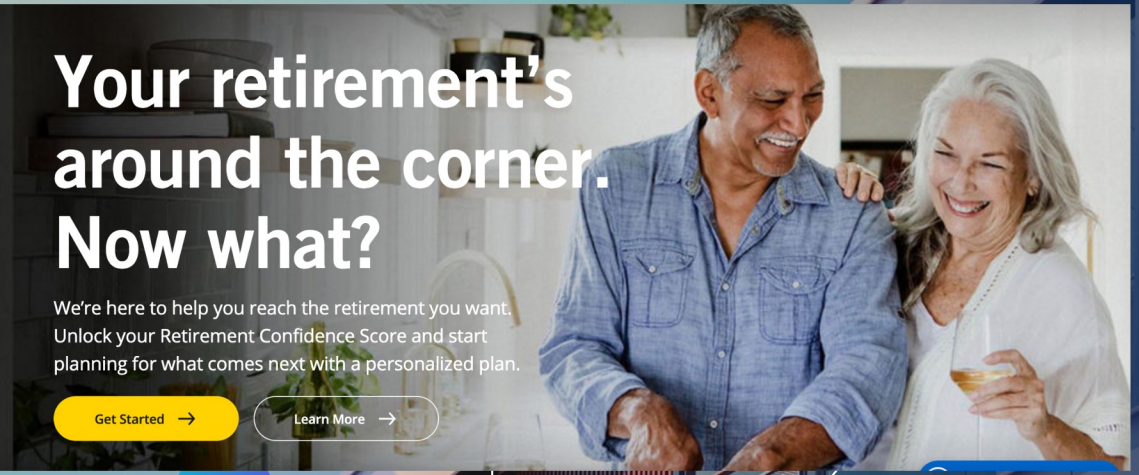
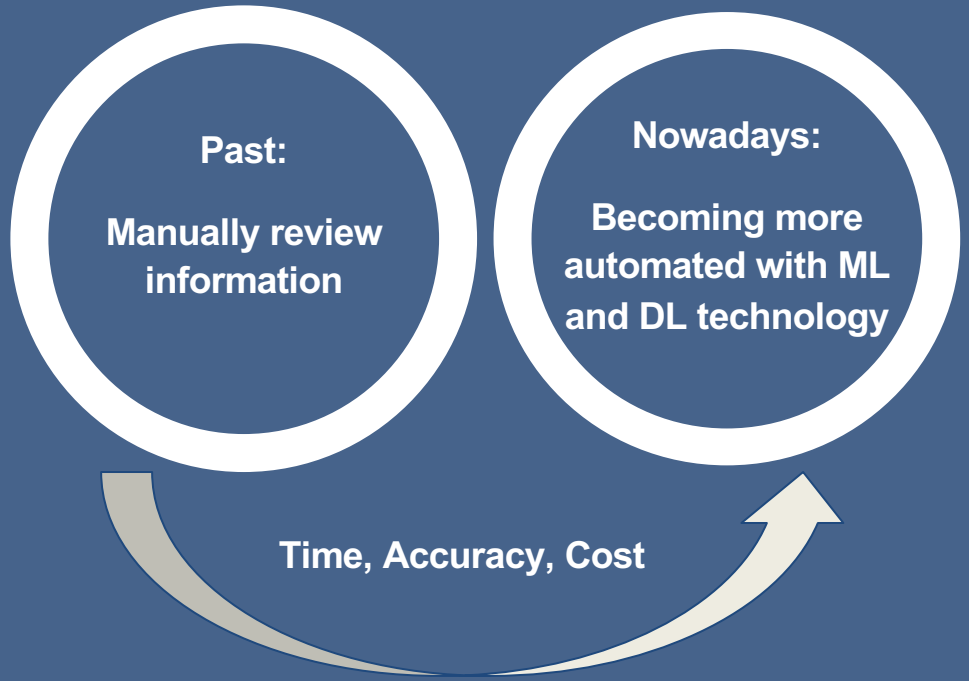
Prudential



Prudential: Life Insurance Assessment

Presenter: Jiaye Li

Business Problem—Pure Risk Premium



MISSION

- Multi-category classification
- Provide supervised ML techniques solutions to risk ratings
- Support underwriters in pricing decision-making process

Data Description

Dataset: The dataset consists of 128 columns with 127 predictors and 1 response variable.
18 numerical variables and 109 categorical variables. Feature name is anonymized.

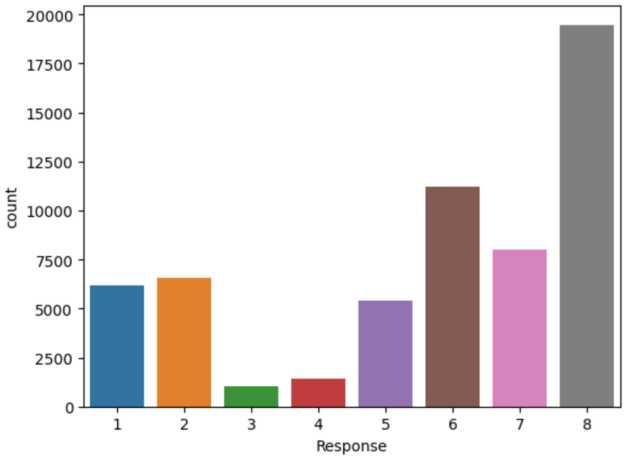
Applicant Info:	Age, Ht, Wt, BMI
Insurance Product Info:	Product_Info_1-7
Employment Info:	Employment_Info_1-6
Insured Info:	InsuredInfo_1-7
Insurance History:	Insurance_History_1-9
Family History:	Family_Hist_1-5
Medical History:	Medical_History_1-41
Medical Keyword:	Medical_keyword_1-48 (dummy encoded)

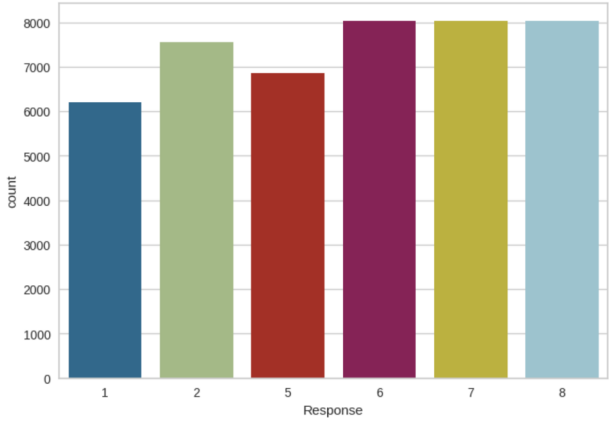
Product_Info_1	Product_Info_2	Product_Info_3	Product_Info_4	Product_Info_5	Product_Info_6	Product_Info_7	Ins_Age	Ht	Wt	...	Medical_Keyword_40	Medical_Keyword_41	Medical_Keyword_42	Med
1	D3	10	0.076923	2	1	1	0.641791	0.581818	0.148536	...	0	0	0	
1	A1	26	0.076923	2	3	1	0.059701	0.600000	0.131799	...	0	0	0	
1	E1	26	0.076923	2	3	1	0.029851	0.745455	0.288703	...	0	0	0	
1	D4	10	0.487179	2	3	1	0.164179	0.672727	0.205021	...	0	0	0	
1	D2	26	0.230769	2	3	1	0.417910	0.654545	0.234310	...	0	0	0	

ws × 127 columns

Exploratory Data Analysis

- The target response variable, Response, is categorical, representing the final decision associated with an application.
- Ordinal scale from 1 to 8.
- Imbalanced dataset, with nearly 20,000 instances in class 8
- Consider undersampling, combine sparse categories

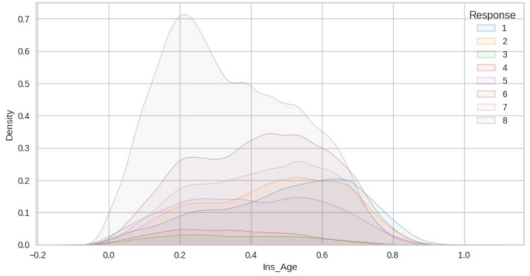


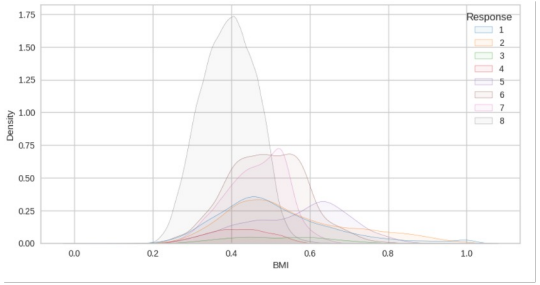


Exploratory Data Analysis

Distribution:

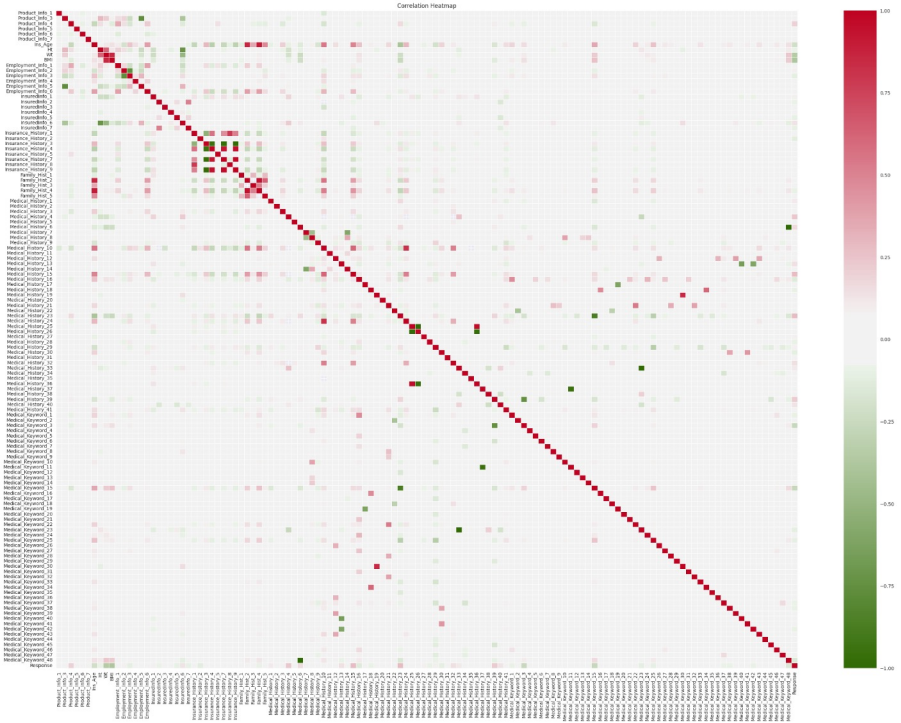
- **Age: younger people have lower risk**
- **BMI: lower BMI indicates lower risk**





Correlation Heatmap:

- **Light correlation with target variable**
- **Height, BMI, Age, Weight , and Medical history 4 are highly correlated**



Data Preprocessing

Missing value:

- Large proportion
- Eliminate entire variable with missing proportion larger than 50%
- Fill with mean value

	missing_count	missing_percent
Employment_Info_1	19	0.03
Employment_Info_4	6777	11.42
Employment_Info_6	10851	18.28
Insurance_History_5	25388	42.78
Family_Hist_2	28644	48.26
Family_Hist_3	34223	57.66
Family_Hist_4	19176	32.31
Family_Hist_5	41789	70.41
Medical_History_1	8885	14.97
Medical_History_10	58794	99.06
Medical_History_15	44570	75.10
Medical_History_24	55553	93.60
Medical_History_32	58244	98.13

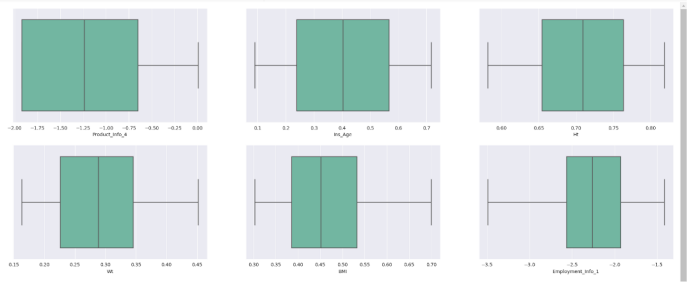
Skewness:

- Light skewness
- Skewness correction

	Skewness
Product_Info_4	1.221772
Ins_Age	0.021939
Ht	-0.086271
Wt	0.693933
BMI	0.943810
Employment_Info_1	4.713368
Employment_Info_4	15.078894
Employment_Info_6	0.787433
Insurance_History_5	102.899632
Family_Hist_2	0.212192
Family_Hist_3	-0.568174
Family_Hist_4	0.206935
Family_Hist_5	-0.727432
Medical_History_1	5.634394
Medical_History_10	-0.309968
Medical_History_15	0.016704
Medical_History_24	1.557523
Medical_History_32	4.342796

Outliers:

- Winsorized



Dummy encoding:

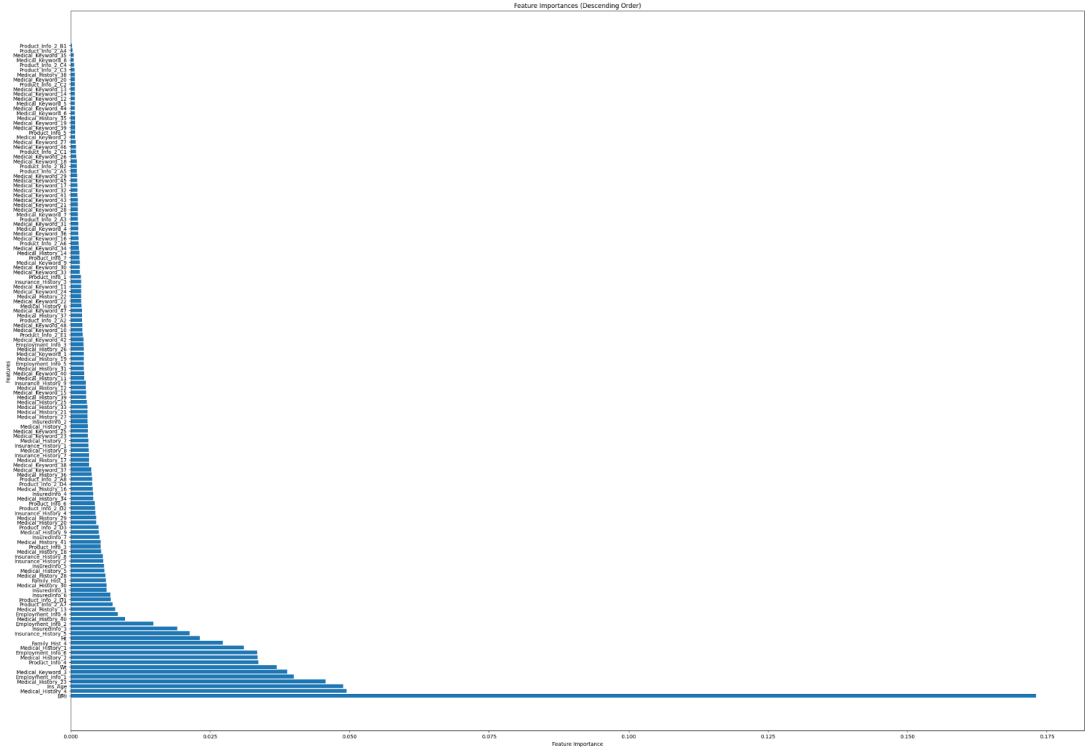
- Product_Info_2

Product_Info_2 Pr
D3
A1
E1
D4
D2

Feature Engineering

Feature importance

- Random Forest Feature Importance
- Select top 35 important features



Holdout Validation

- 80-20 training vs test set split

```
Shape of X and y (59351, 35) (59351,)
Shape of training set X and y (47480, 35) (47480,)
Shape of test set X and y (11871, 35) (11871,)
```

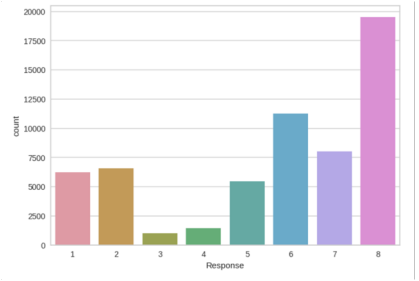
Scaling

- Minmax scaler

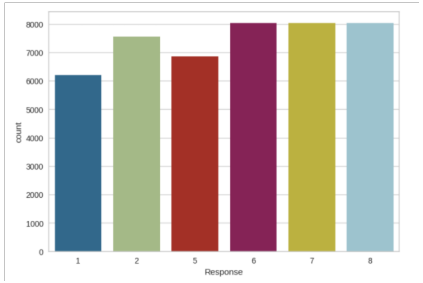
Label Encoding

- Multi-category classification
- Transform target variable from 1,2,3,..8, to 0,1,2,...,7

Model Development



8 Classes



5 Classes(Undersampling)

Benchmark (PyCaret)

	Model	Accuracy
lightgbm	Light Gradient Boosting Machine	0.5345
gbc	Gradient Boosting Classifier	0.5284
rf	Random Forest Classifier	0.5172
et	Extra Trees Classifier	0.5141
lda	Linear Discriminant Analysis	0.4922
ada	Ada Boost Classifier	0.4891

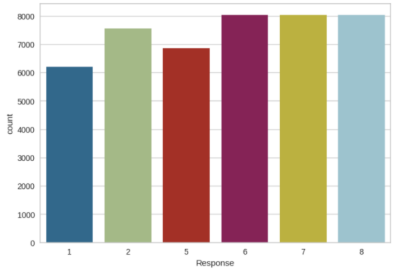
	Model	Accuracy
lightgbm	Light Gradient Boosting Machine	0.4598
gbc	Gradient Boosting Classifier	0.4505
rf	Random Forest Classifier	0.4421
et	Extra Trees Classifier	0.4351
lda	Linear Discriminant Analysis	0.4133
ada	Ada Boost Classifier	0.4167

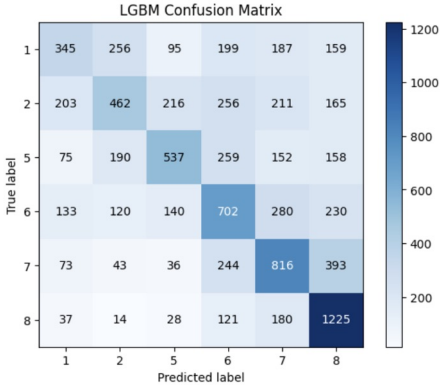
Model (5 Classes)

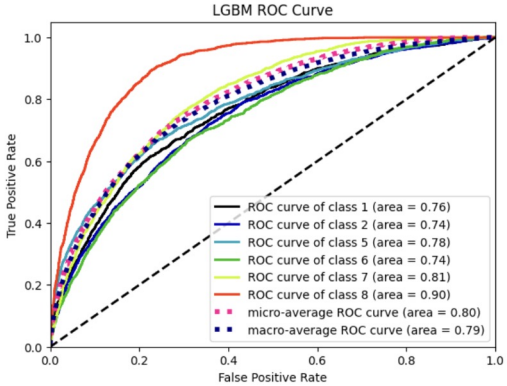
Classifier	Accuracy
Light GBM(softmax)	0.446
XGBoost	0.442
MLP	0.426
AdaBoost	0.416
Decision Tree	0.402
Random forest	0.392
SVM(rbf)	0.329
SVM(poly)	0.257

Classifier	Accuracy
Light GBM	0.452
Stacked Classifier	0.449
XGBoost	0.448

```
StackingClassifier
  RF: RandomForestClassifier
  LGB: LGBMClassifier
  final_estimator: XGBClassifier
```







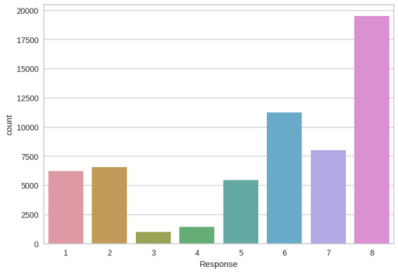
Model (8 Classes)

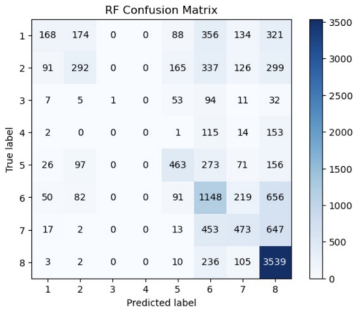
Classifier	Accuracy
OVO LGB Classifier	0.5261
LGB Stacked (ova)	0.5261
XGB Stacked	0.5256
Light GBM (multiclass)	0.5250
Light GBM (ova)	0.5249
RF Stacked	0.5246
MLP Stacked	0.5245
XGBoost	0.5240
MLP	0.5153
Random forest	0.5094

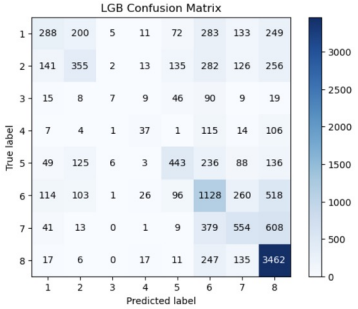
One-Vs-All & One-Vs-One

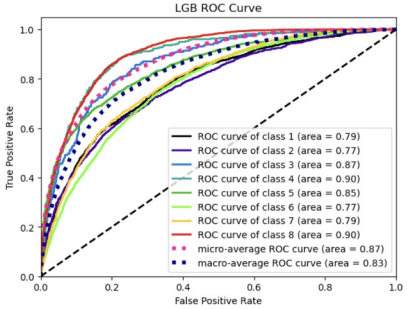
red vs [blue, green]
blue vs [red, green]
green vs [red, blue]

red vs. blue
red vs. green
red vs. yellow
blue vs. green
blue vs. yellow
green vs. yellow

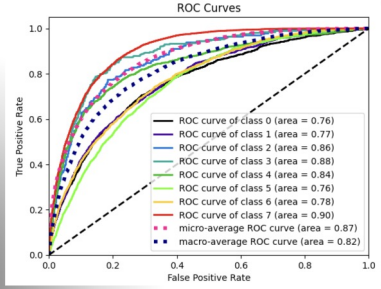


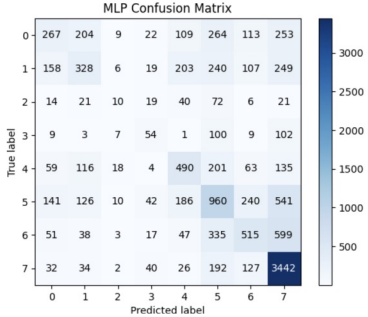


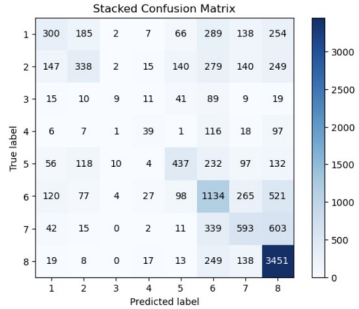


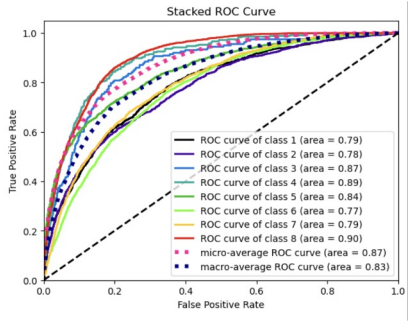


Keras MLP: 0.5160



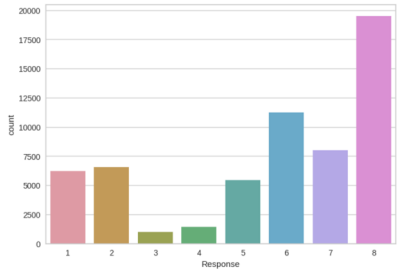




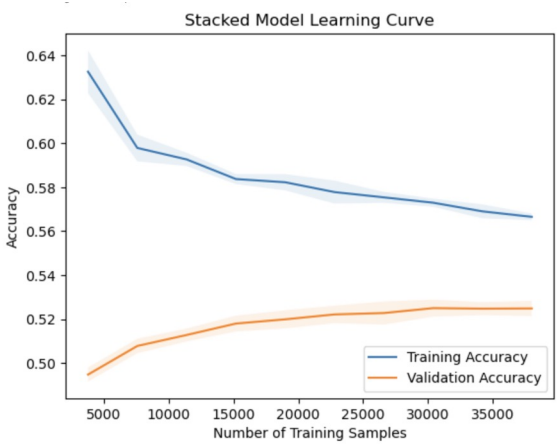


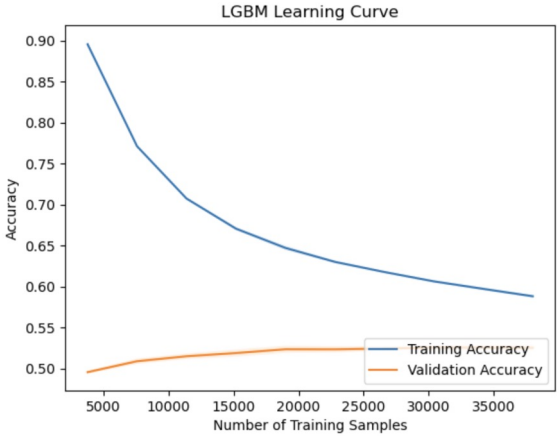
Model (8 Classes)

Classifier	Accuracy
OVO LGB Classifier	0.5261
LGB Stacked (ova)	0.5261
XGB Stacked	0.5256
Light GBM (multiclass)	0.5250
Light GBM (ova)	0.5249
RF Stacked	0.5246
MLP Stacked	0.5245
XGBoost	0.5240
MLP	0.5153
Random forest	0.5094



Fairly Fitted to the Data

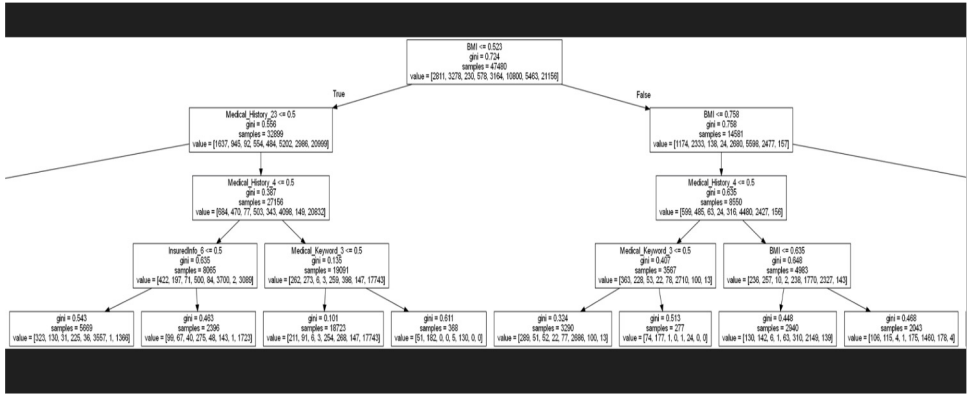




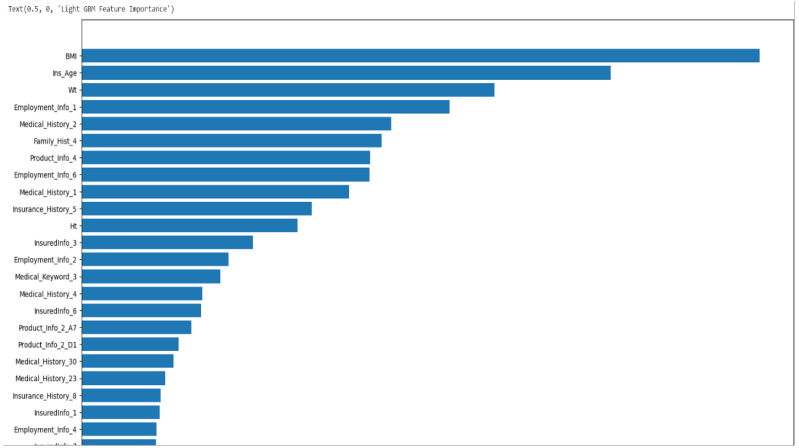
Model Explanation (XAI–LightGBM)

Surrogate Model : Decision Tree

- **Depth =4**
- **Select four features:**

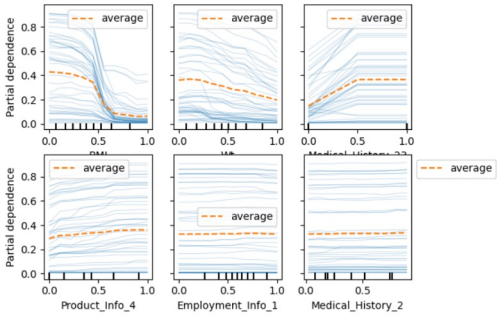


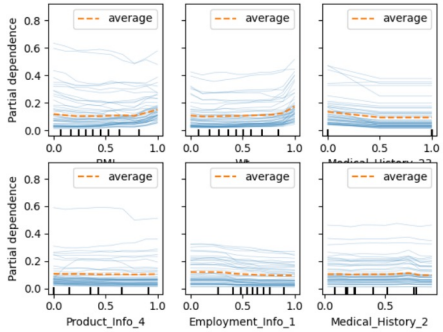
Feature importance



Partial Dependence

- **Compare target =1 (R) with target =8(L)**
- **Similar outcome with feature importance and EDA**



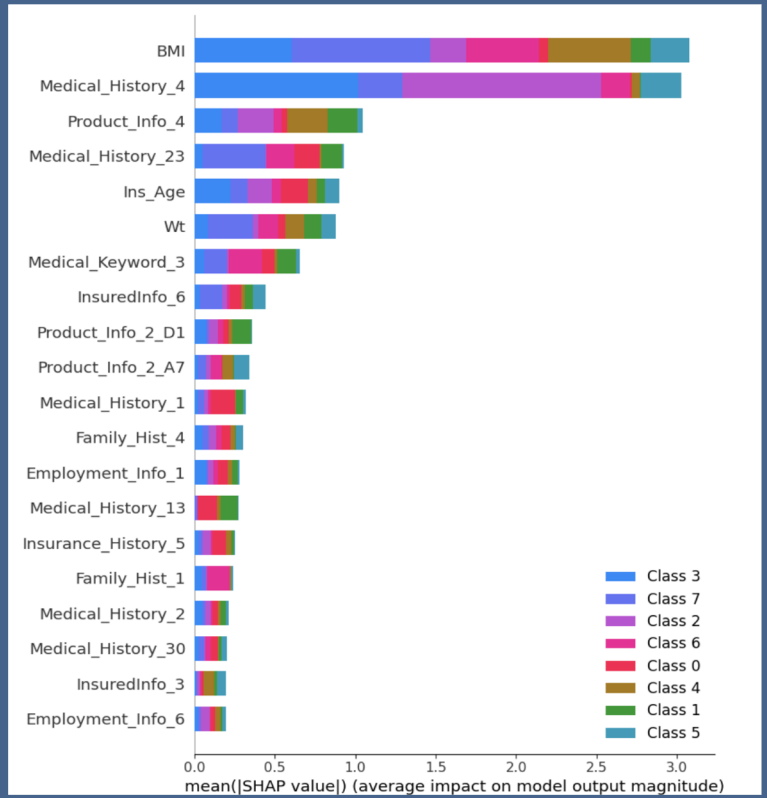


Conclusion

- The model is generalized to predict across the entire distribution to a reasonable degree of accuracy, with highest value belonging to class 8.

- Some features play a significant role in decision making process, for example, BMI, Medical_History_23, and Product_Info_4.

- The model could be more useful if assigned to predict fewer risk classes, aiding decision process.



SWOT in Deployment

STRENGTHS

- Reduce Model Complexity: Reduce lots of features only at expense of less than 1% accuracy.
- Interpretable Model: Explain feature importance and their impact to classification.
- Proximity of the misclassification cases: class 8 is classified as class 7, few cases of class 8 classified as class 1

WEAKNESSES

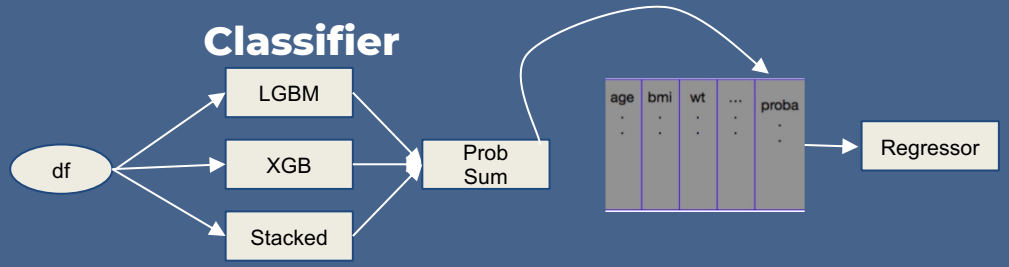
- Limitation in Accuracy
- Low true positive rate in minority classes (class 3 & 4)

OPPORTUNITIES

- Ensemble model & Regression Problem (Round up method)
  - Lose accuracy, but increase proximity
- Other Policy Mechanism
- Monitor and Update Model with new data

THREATS

- Data Privacy and Security: handling sensitive customer data, making data privacy and security paramount. Any breaches or mishandling of data lead to reputational damage and legal consequences.



Prudential

Your retirement's around the corner. Now what?

We're here to help you reach the retirement you want. Unlock your Retirement Confidence Score and start planning for what comes next with a personalized plan.

Get Started →

Learn More →

Thank You