2022

PREDICTIVE UNDERWRITING RISK SCORING MODEL

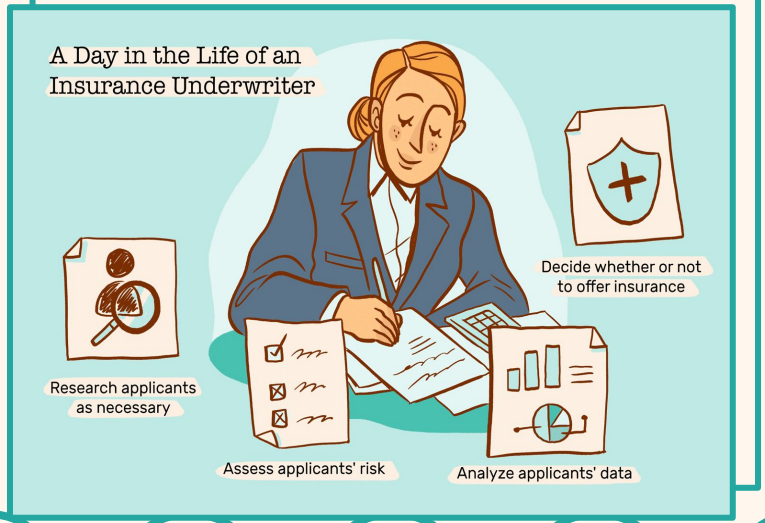
By: Marc Tan

Oct 2022

TABLE OF CONTENTS

01 DATASET

- Prudential Life Insurance Assessment (Kaggle)
- Roughly 80,000 rows & 128 columns
- Masked / Anonymised
- Risk scoring : Accept / Decline
- Contains typical new applicant information for Life Insurance.
- Product info, Applicant info, Employment info, Insurance, Medical & Family history, etc



BACKGROUND 01

- At present, Underwriters review a high volume of Life Insurance applications manually on a daily basis.
- Traditionally, Underwriters consult an actuarial table to see the mortality probability of the applicant. The higher the probability, the higher your premium.
- This results in inefficiencies, inaccuracies and overall longer turnaround times before policy is issued or offered.

01

PROBLEM STATEMENT

The conventional Underwriting process is:

- Costly (Skilled labour, Medical Reports)
- Time consuming & slow (Follow-ups, Manual)
- Inconsistent (different underwriters, different opinions)

but a necessary exercise for direct life insurance underwriters.

Inefficiencies

Inaccuracies

Expensive

Tired

Efficient

Cost savings

Fair & Ethical

SOLUTION

Develop a supervised machine learning model to predict the underwriting decision for new applicants based on the data completed in their application form.

LIFE INSURANCE APPLICATION FORM

10 HEALTH DETAILS OF PROPOSED INSURED – To be completed for non-medical application, or where the medical examination was done earlier than the application form signed date.

10.1 a. Height (metres):

b. Weight (kilograms):

c. Was there any weight change in the past year? ◯ Yes ◯ No
If yes, how much and state the reason:

d. Name and Address of the Proposed Insured's Regular Doctor:

e. When did you last consult a doctor? Please provide reason, name of clinic (if differs from 10.1.d) and result of the last consultation:

10.2 Have you ever used any habit forming drugs or narcotics or been treated for drug habits or consumed alcohol excessively or been treated for alcoholism? ◯ Yes ◯ No

10.3 Have you ever had or been told to have or been treated for:

a. epilepsy, fits, stroke, paralysis, weakness of limb, prolonged headache, unconsciousness, nervous breakdown, depression or any other nervous/mental disorders? ◯ Yes ◯ No

b. diabetes, thyroid disorders or any other endocrine disorders? ◯ Yes ◯ No

c. ear discharge, nose bleeds, double vision, impaired sight, hearing, or speech or any other disorders of ear, eye, nose or throat? ◯ Yes ◯ No

d. asthma, persistent cough, coughing with blood, pneumonia, tuberculosis, chest or breathing complaints/ discomfort or any other lung disorders? ◯ Yes ◯ No

e. raised cholesterol, high blood pressure, heart attack, heart murmur, cardiomyopathy, mitral valve prolapse or other heart valve disorders, breathlessness, irregular or fast heart rate, chest discomfort or pain, disease of or any other disorders of the heart or blood vessels? ◯ Yes ◯ No

f. gastritis, stomach or duodenal ulcer, blood in stools, fistula, piles or any other stomach or bowel disorders? ◯ Yes ◯ No

g. jaundice, hepatitis B carrier or any form of hepatitis, liver disorder or gall bladder disorder? ◯ Yes ◯ No

h. blood, protein or sugar in urine, kidney stones, infection or any other disorders of the kidney, bladder or genital organs? ◯ Yes ◯ No

i. slipped disc, gout, arthritis, pain or deformity or disorders of the muscles, spine, limbs or joints or severe injury? ◯ Yes ◯ No

j. cancer, tumours, cysts or growths of any kind? ◯ Yes ◯ No

k. anaemia, any other disorders of the blood, advised to abstain from donating blood or received blood transfusion or blood products on account of haemophilia or any other reason? ◯ Yes ◯ No

l. any other illness, disorder, operation, physical disability or accident not mentioned above? ◯ Yes ◯ No

10.4 Are you awaiting or intending to have any medical consultations, investigations or treatment; or experiencing any symptoms that might cause you to seek medical treatment in the near future? ◯ Yes ◯ No

10.5 Have you or your spouse been told to have, received any medical advice, counselling or treatment in connection with sexually transmitted disease, AIDS, AIDS Related Complex or any other AIDS related condition? ◯ Yes ◯ No

10.6 a. Have you ever had HIV testing done? ◯ Yes ◯ No
If yes, please state reason, date and results:

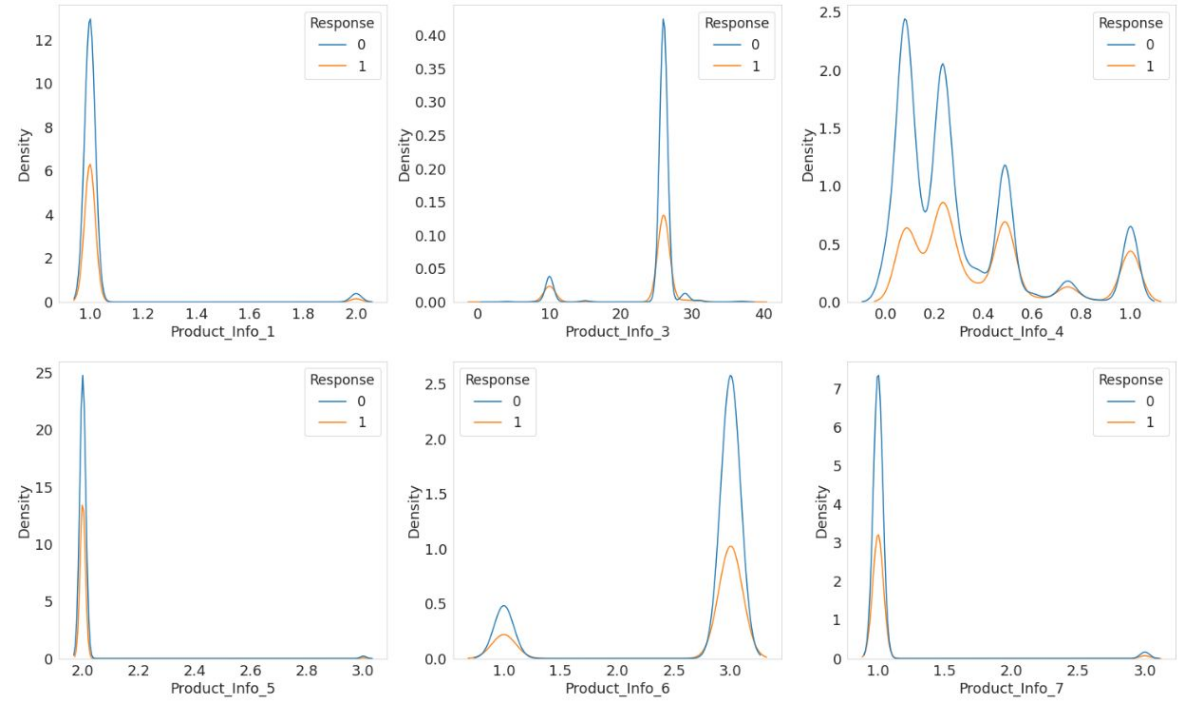
b. In the last 3 months have you had any of the following symptoms for more than one week continuously: fatigue, weight loss, diarrhoea, enlarged nodes or unusual skin lesions? ◯ Yes ◯ No
If yes, please state reason, date and results:

EDA - GROUPING OF FEATURES

```
Group_ProdInfo = ['Product_Info_1','Product_Info_2','Product_Info_3',
                  'Product_Info_4','Product_Info_5','Product_Info_6',
                  'Product_Info_7']

Group_ApplicantInfo = ['Ins_Age','Ht','Wt','BMI']

Group_EmploymentInfo = ['Employment_Info_1','Employment_Info_2','Employment_Info_3',
                        'Employment_Info_4','Employment_Info_5','Employment_Info_6']

Group_InsuredInfo = ['InsuredInfo_1','InsuredInfo_2','InsuredInfo_3',
                     'InsuredInfo_4','InsuredInfo_5','InsuredInfo_6',
                     'InsuredInfo_7']

Group_InsuranceHistoryInfo = ['Insurance_History_1','Insurance_History_2','Insurance_History_3'
                              ,'Insurance_History_4','Insurance_History_5','Insurance_History_7',
                              'Insurance_History_8','Insurance_History_9']

Group_FamilyHistoryInfo = ['Family_Hist_1','Family_Hist_2','Family_Hist_3',
                           'Family_Hist_4','Family_Hist_5']

Group_MedicalHistoryInfo = ['Medical_History_1','Medical_History_2','Medical_History_3',
                            'Medical_History_4','Medical_History_5','Medical_History_6',
                            'Medical_History_7','Medical_History_8','Medical_History_9',
                            'Medical_History_10','Medical_History_11','Medical_History_12',
                            'Medical_History_13','Medical_History_14','Medical_History_15',
                            'Medical_History_16','Medical_History_17','Medical_History_18',
                            'Medical_History_19','Medical_History_20','Medical_History_21',
                            'Medical_History_22','Medical_History_23','Medical_History_24',
                            'Medical_History_25','Medical_History_26','Medical_History_27',
                            'Medical_History_28','Medical_History_29','Medical_History_30',
                            'Medical_History_31','Medical_History_32','Medical_History_33',
                            'Medical_History_34','Medical_History_35','Medical_History_36',
                            'Medical_History_37','Medical_History_38','Medical_History_39',
                            'Medical_History_40','Medical_History_41']
```

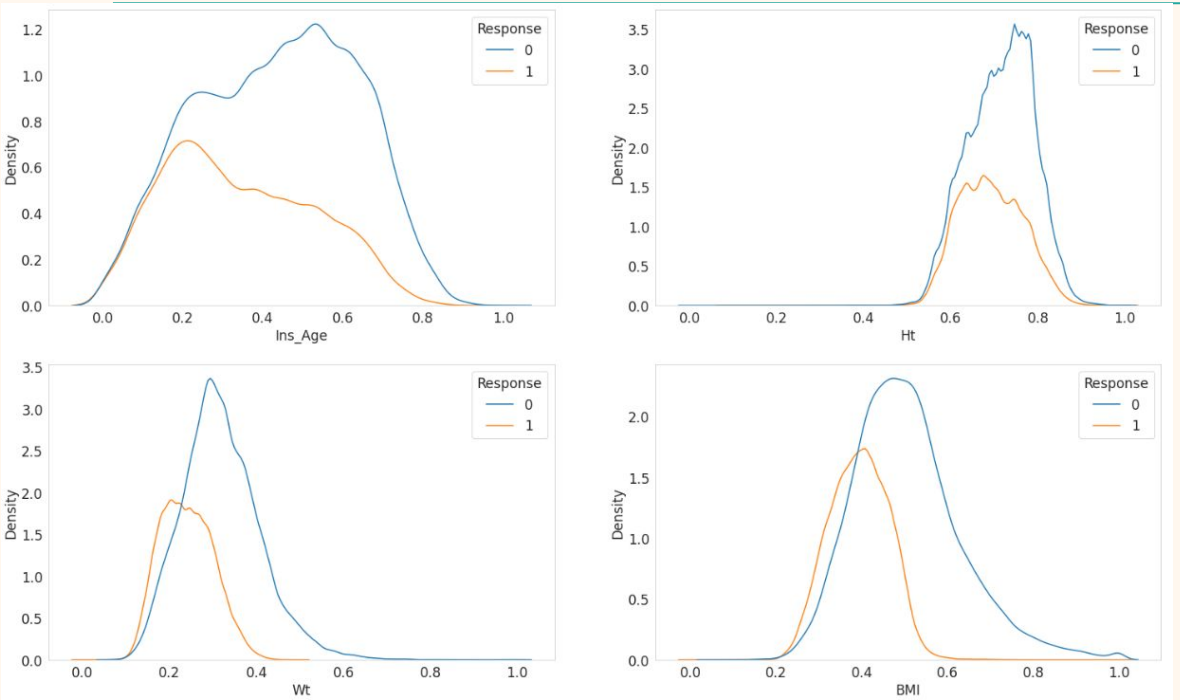
- Grouped features into a more general category.
  - Product info
  - Applicant info
  - Employment info
  - Insured info
  - Insurance history info
  - Family history info
  - Medical history info
- Generate KDE plots for each feature. Target response (Accept/Decline) differentiated by hue.
- Helps better understand if there are any trends/correlations within the data

EDA - PRODUCT INFO



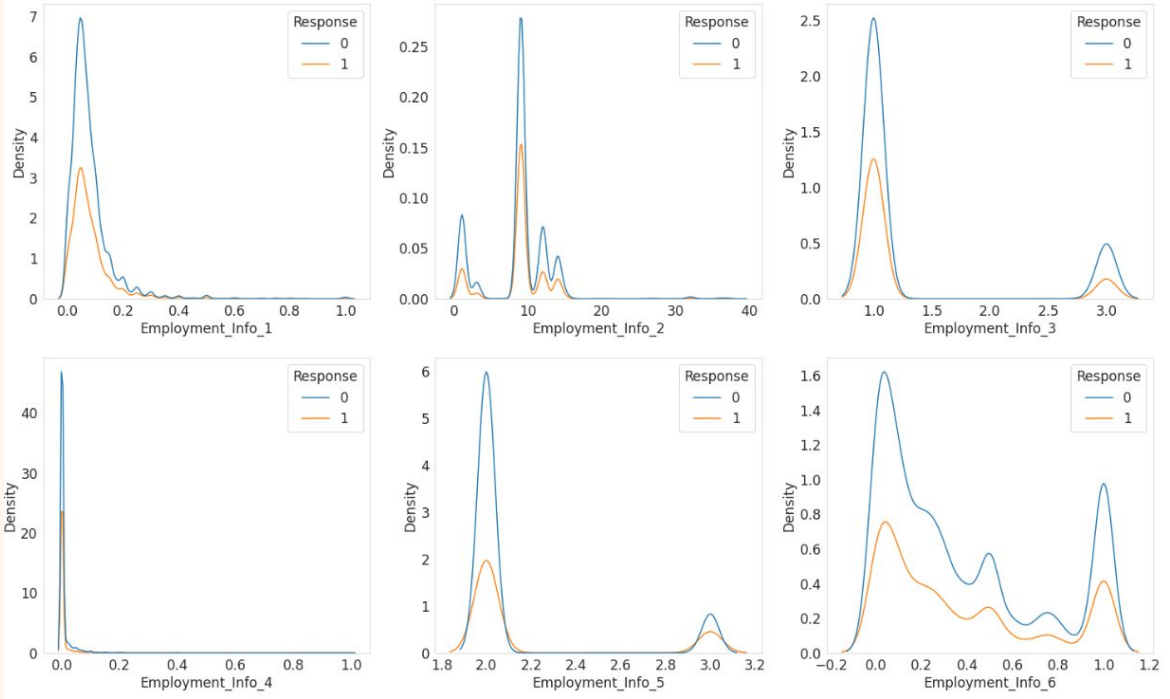
- Consistent trend to features
- Closely overlap between each response class
- No major difference in relative densities
- Unlikely to affect or help towards predicting applicants' risk rating

EDA - APPLICANT INFO



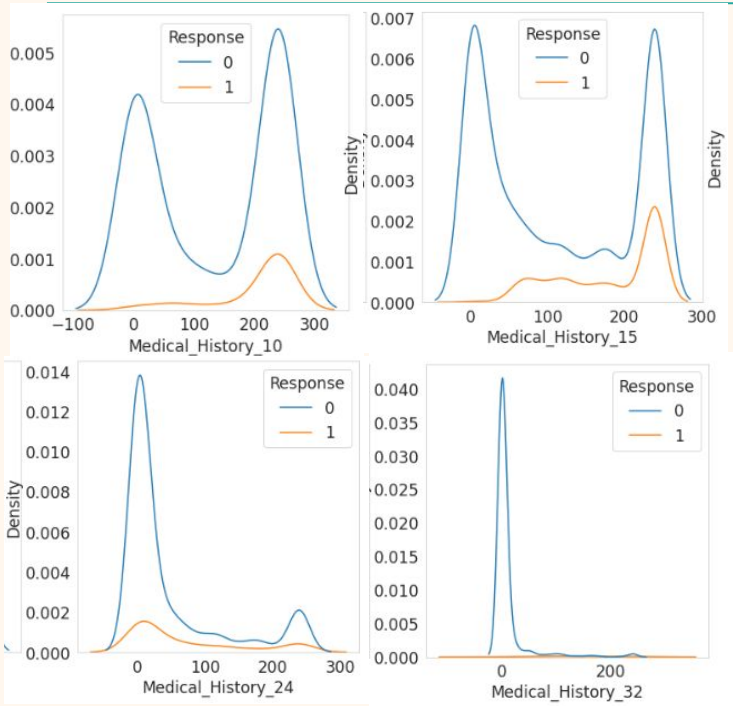
- Variation in each response class distribution
- Positive vs negative skew
- Spread for each feature differs
- Applicant info plays some part in correlation of our target response

EDA - EMPLOYMENT INFO



- Similar to product info covered earlier.
- Consistent trend to features
- Closely overlap between each response class
- Unlikely to affect or help towards predicting applicants' risk rating

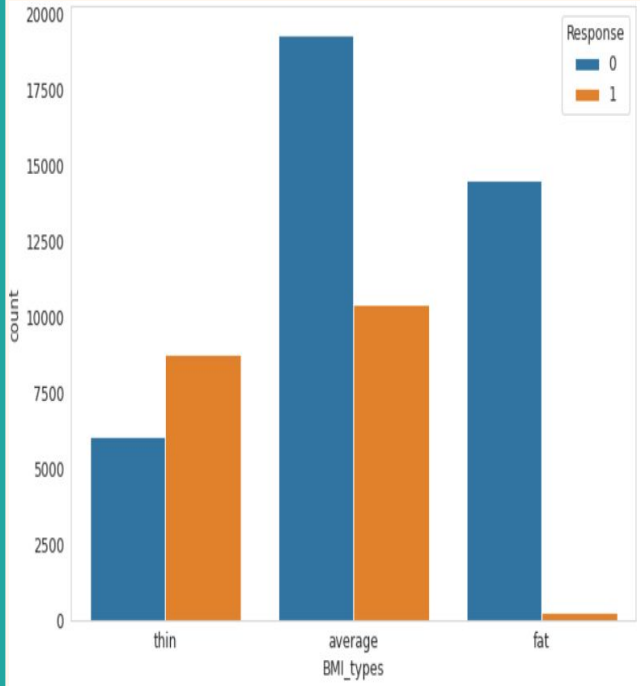
EDA - MEDICAL HISTORY

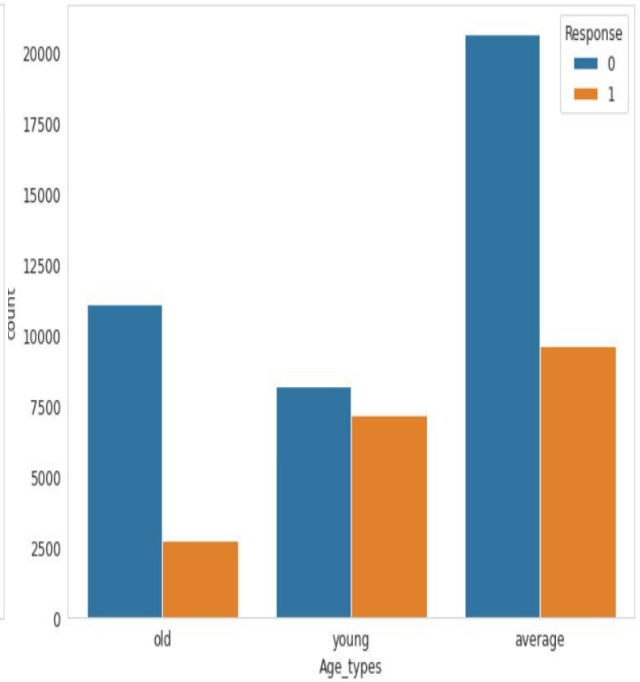


- Like applicant info, these KDE features some degree of predictive distinction in terms of variance.
- However, worth noting that this consist of a high proportion of missing values.
- Distribution should not be misconstructed.

SHAPE DIFFERENCES

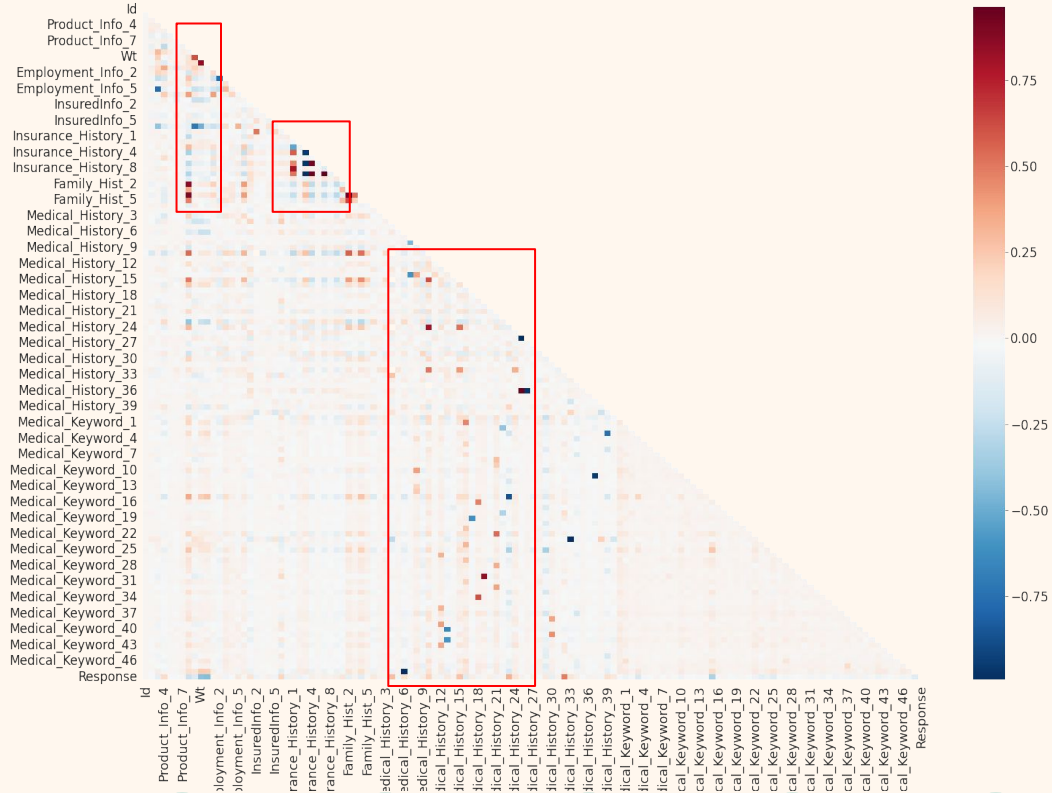
EDA - BMI & AGE





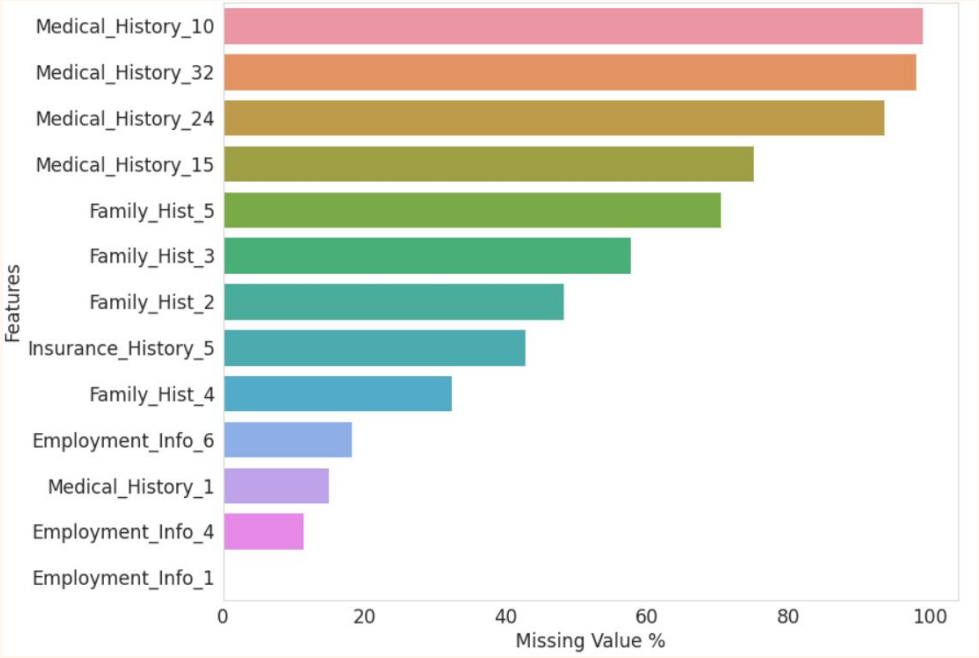
- Categorized by quantile for BMI & AGE
- "Fat" & "Old" applicants less likely to get offered cover
- Logical as obesity and older ages comes with increased risk factors

CORRELATION HEATMAP



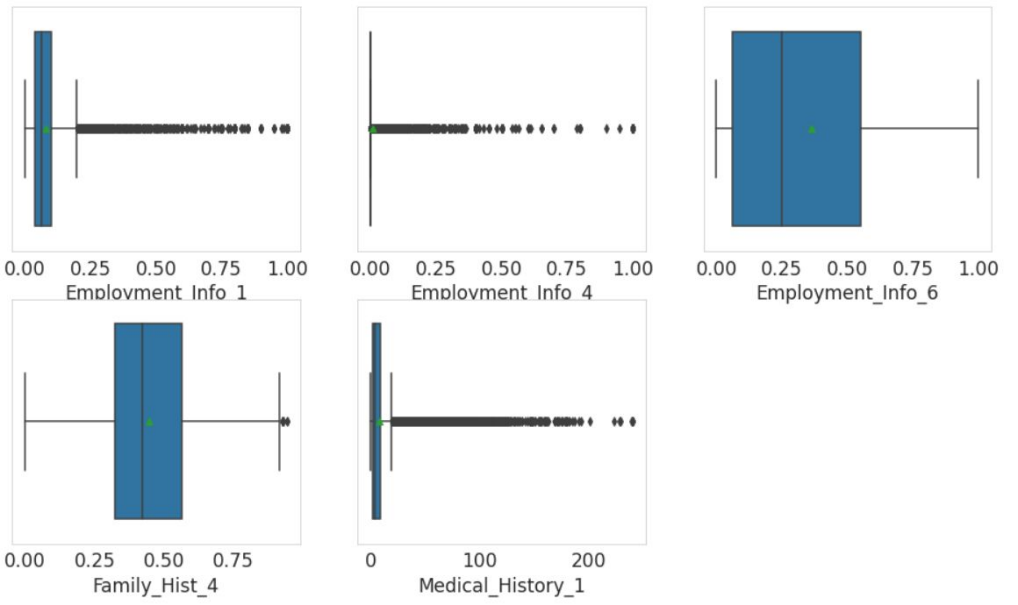
- Strong correlations noted for Applicant info portion and Employment, Insured Hx and Fhx (Lifestyle - think career with age, Genetics & FHx)
- Insurance History : Strong correlations with other Insurance_history columns and some Fhx. (Do you have an existing policy? & How much is your existing cover)
- Medical History : Some columns show a number of correlation hotspots against several Medical_History & Medical_Keyword columns but do not show any notable interactions with the rest of the features.

DATA CLEANING



- Explore missing values from dataset
- Some medical histories missing
- Consideration to fill missing values
- Cleared features with missing values in excess of 40%

DATA CLEANING



- Remaining features with missing values <40%
- Positive skewed / Outliers present
- Fill missing values with median.

MODELLING & EVALUATION

01 GRID SEARCH CROSS-VALIDATION TECHNIQUE

02 PRECISION - RECALL CURVE

03 SHAP & FEATURE IMPORTANCE

MODEL RESULTS

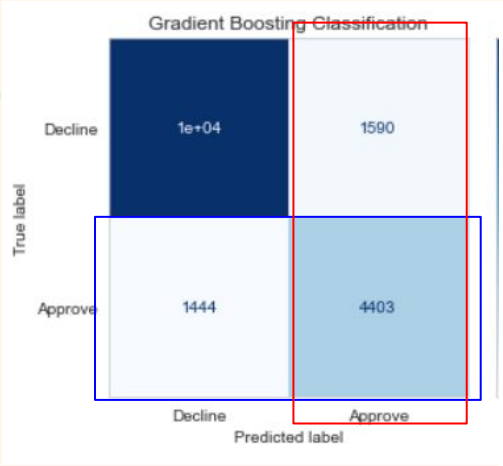
	Train score	Test score	Generalisation	Accuracy	Precision	Recall	Specificity	F1	ROC AUC
Logistic Regression	0.793	0.796	-0.378	0.796	0.644	0.849	0.770	0.732	0.8825
k-Nearest Neighbour Classification	1.000	0.742	25.800	0.742	0.600	0.644	0.790	0.621	0.7981
Random Forest Classification	0.809	0.805	0.494	0.805	0.752	0.604	0.903	0.670	0.8859
Decision Tree Classification	0.820	0.809	1.341	0.809	0.688	0.766	0.830	0.725	0.8795
XGBoost Classification	0.827	0.824	0.363	0.824	0.718	0.762	0.854	0.739	0.8991
AdaBoost Classification	0.828	0.824	0.483	0.824	0.717	0.769	0.852	0.742	0.9003
Gradient Boosting Classification	0.837	0.830	0.836	0.830	0.736	0.754	0.868	0.745	0.9051

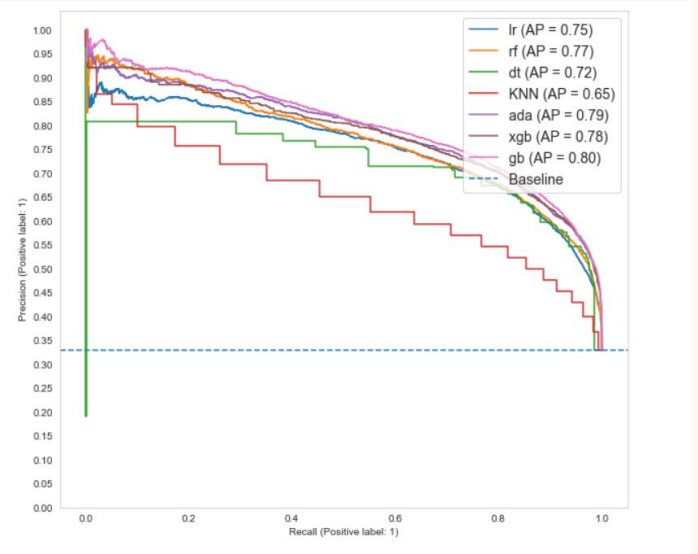
1. 7 models performed via GridSearchCV
2. Chosen model: Gradient Boosting model
   a. Highest F1 score
   b. Generalisation <5%
   c. Highest accuracy

Model	GB Classifer
Transformer	MinMax Scaler
Learning rate	0.1
n_estimators	100
max_depth	5
min_sample_leaf	2
min_samples_split	5

MODEL RESULTS

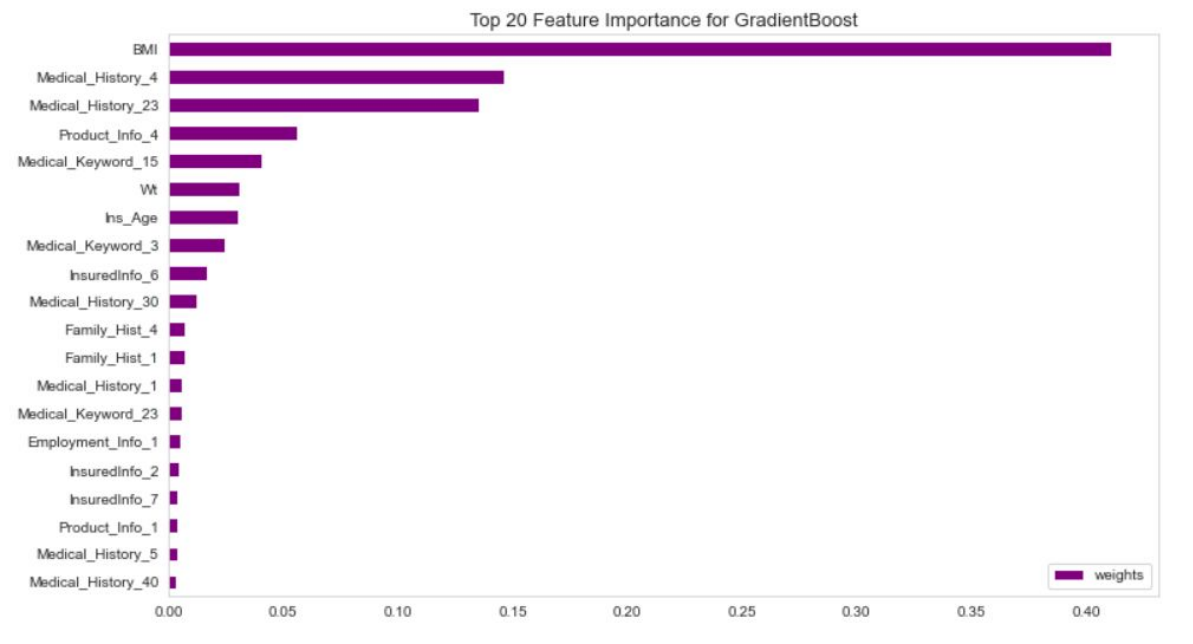
- Problem statement: Develop a supervised machine learning model to predict the underwriting decision for new applicants based on the data completed in their application form.
- Focus area on highest True Positive label
- Business considerations for FP & FN
- FP results in potential claims on “sub-standard lives”
- FN results in potential missed opportunities / business



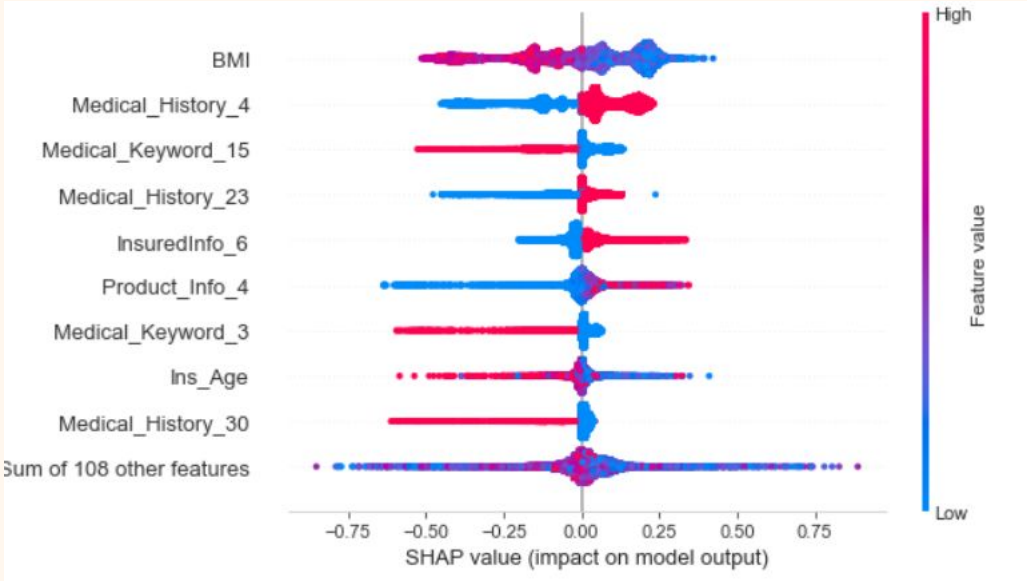


PR CURVE

FEATURE IMPORTANCE



SHAP



- Notably, higher values of BMI & Ins_Age have a negative impact on the prediction while lower values have a positive impact.
- Some other well differentiated medical histories / keywords also noted.

CONCLUSION

- From our results, we have selected the Gradient Boosting Classifer model. The model has the highest Accuracy & AUC PR curve score. Indicating the best performance amongst the other models for our specific problem statement.
- As the datasets feature names have been anonymised, it has not been possible to incorporate "business logic" but rather on a best effort basis. With the help of a "deanonymised" dataset as well as the input of subject matter expertise, it may be possible to create more predictive features and validate them for use in production.
- Albeit limited EDA can be performed due to masked data, the top features are consistent with real world application such as applicant profile (BMI , AGE , Medical Hx).
- FP & FN cases are to be considered on a risk-to-reward basis. Different stakeholders to consider. (Cost savings vs lost business and potential losses).

RECOMMENDATIONS

- Consider using Stacking techniques to compare the result gains by stacking the top 3 models.
- In order to mitigate the risk, may wish to consider limiting model predictions to certain products or by face value (base mortality, low sum assured cover).
- Fine tuning model for precision performance by increasing threshold such that this model may be implemented in the STP process. Cost savings would be quantified by STP cases not requiring review. Rejected applications will be reviewed by existing underwriters and thus no extra implications to resources.

THANK YOU

Do you have any question?

Special thanks to Justin, Jeryll , Stephen and classmates for the enjoyable learning journey!