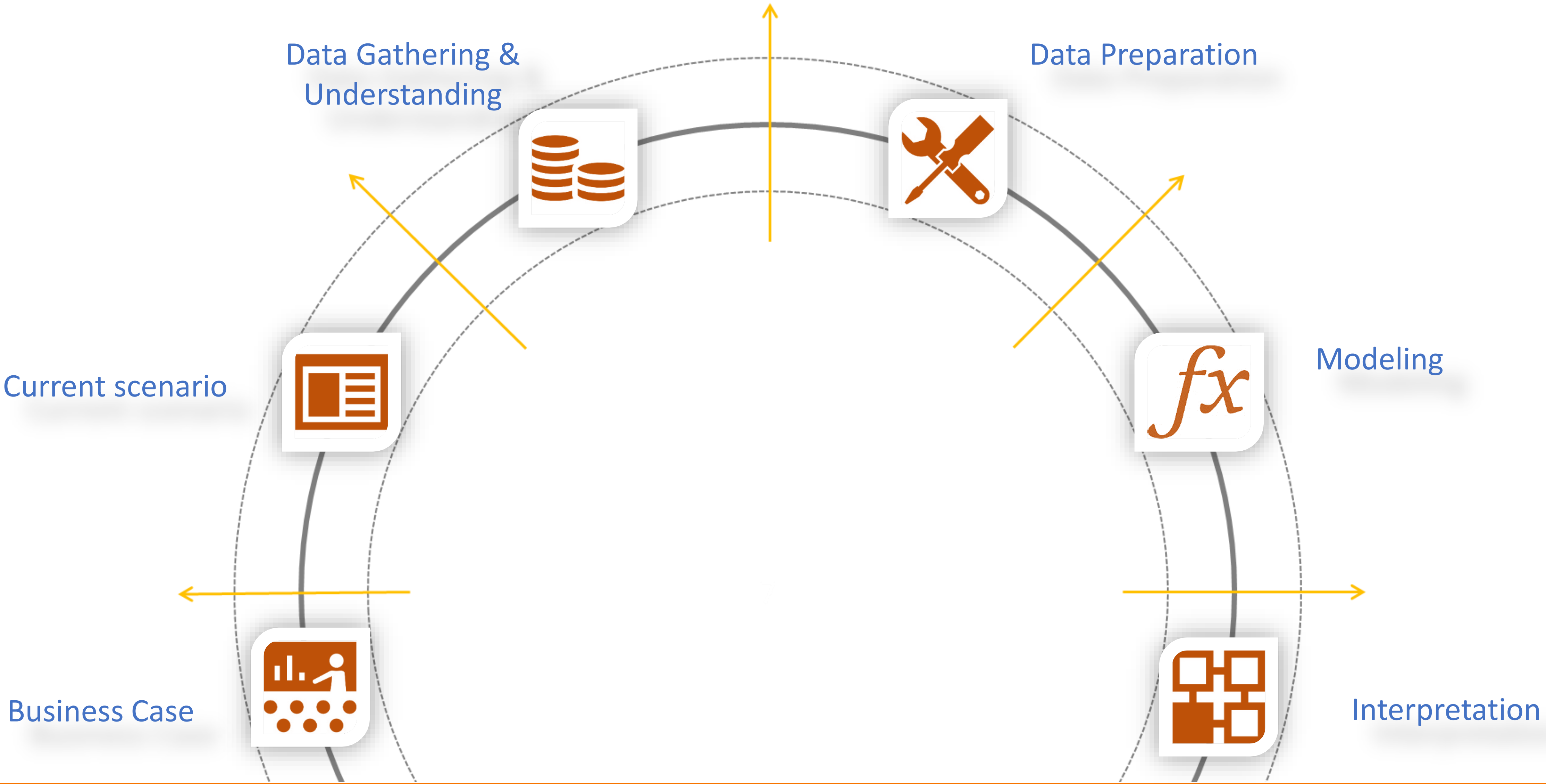
Harnessing the Power of Data Science in Policy Underwriting

Ashok Shetty

MBA in Business Analytics,

REVA University, Bengaluru

Agenda



Data Gathering &
Understanding
Data Preparation
Current scenario
Modeling
Business Case
Interpretation

Problem statement - *The Bigger Picture*

REVA UNIVERSITY

Problem statement

- Insurance industry has not been quick to embrace the digital revolution
- Digital transformation in life insurance is still in the beginning stage
- Life and health insurance products have not changed much since their inception in the 1960s.
- Life and health insurance is still relying on -
  - old-fashioned broker to client interactions
  - lot of face to face communications and paper works
- The overall market penetration of life and health has been declining for the last 30 years and the annual sales of new insurance policies have declined from about 17 million contract in the 1980's to less than 10 million today

Changing landscape of Life & Health Life Insurance Industry

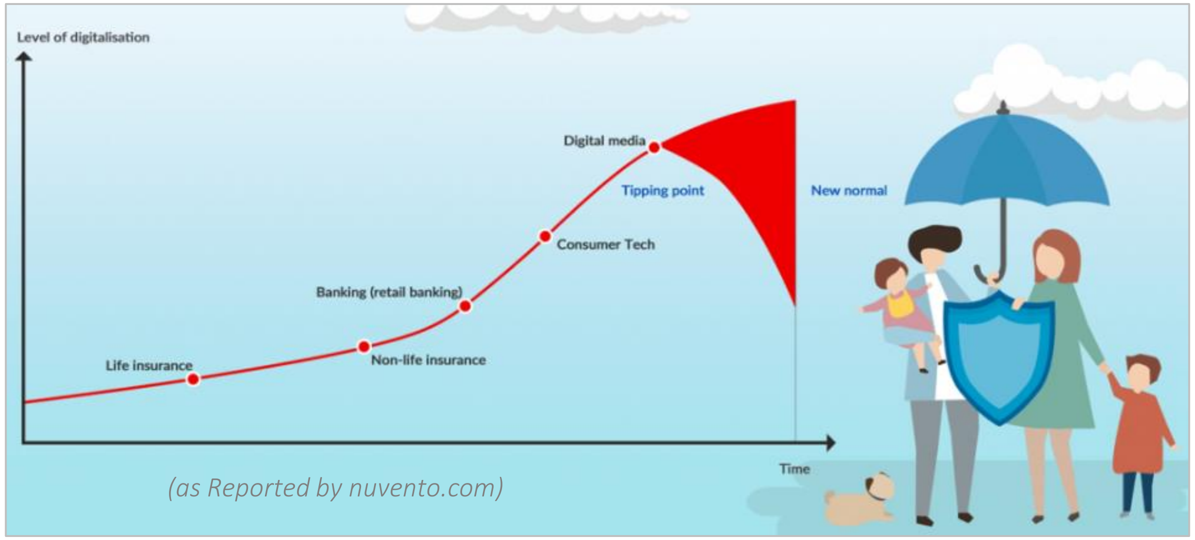
- Due to the pandemic, migrating old–fashioned business to a digital platform has become need of the hour
- Modern consumers are accustomed to speed, transparency and convenience regardless of the channel they use or product they purchase. Insurance customers are no different
- 70% of customers looking for life insurance policies begin their information gathering process online
- Websites of top 10 life insurance providers get more than 7 million total visits a month

The need for a Digital Solution

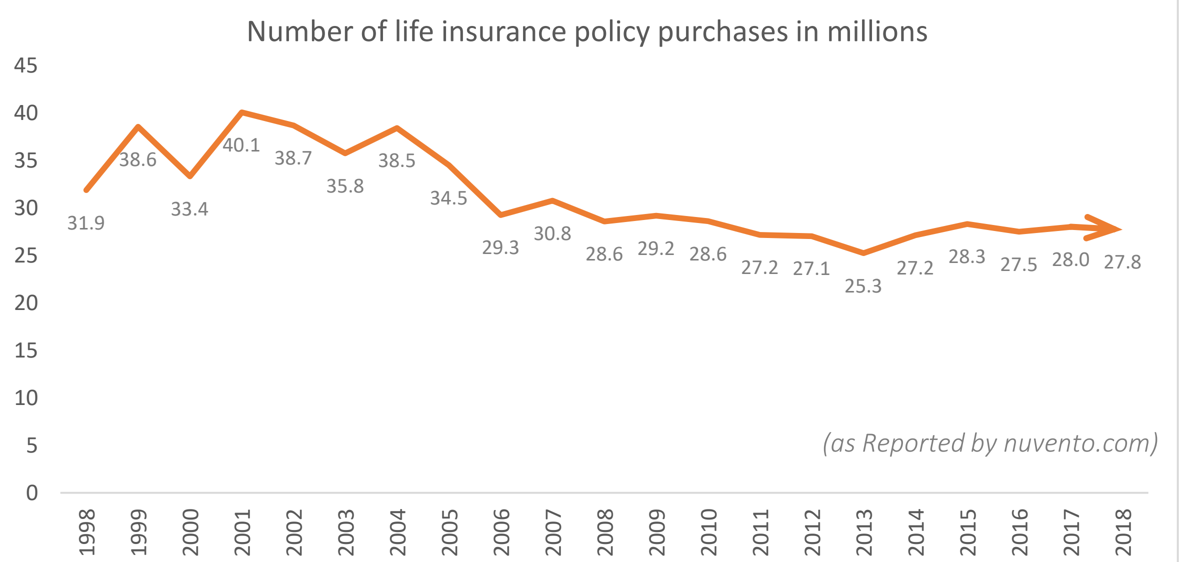
- An online mechanism that can automatically classify the life insurance applications into appropriate risk buckets with highest possible accuracy
- A predictive model that can reduce the 'from application to policy purchase' time by a great extent, thereby increasing the sales and customer satisfaction

Potential benefits

- An online mechanism that can automatically classify the life insurance applications into appropriate risk buckets with highest possible accuracy.
- A predictive model that can reduce the 'from application to policy purchase' time by a great extent, thereby increasing the sales and customer satisfaction



(as Reported by nuvento.com)



(as Reported by nuvento.com)

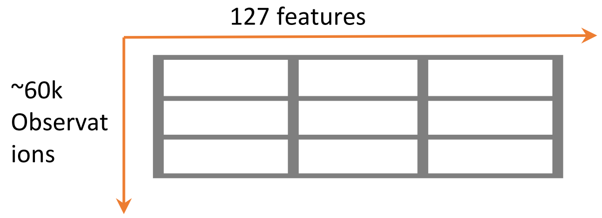
Understanding Data – The Structure

REVA UNIVERSITY

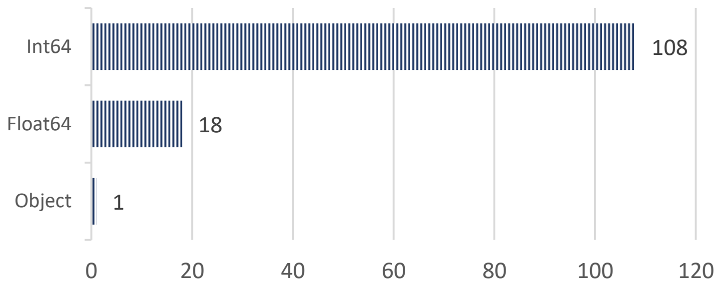
Highlights

- The data we have used is a collection of policy level information of random customers.
- Each observation has been classified on three risk categories by the underwriters.
- The data broadly contain the following feature-sets: Body structure, Product, Employment Information, Insurance History, Family History and Medical History

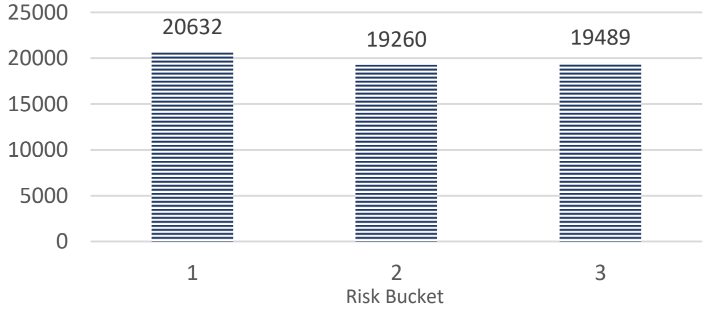
Data Dimension



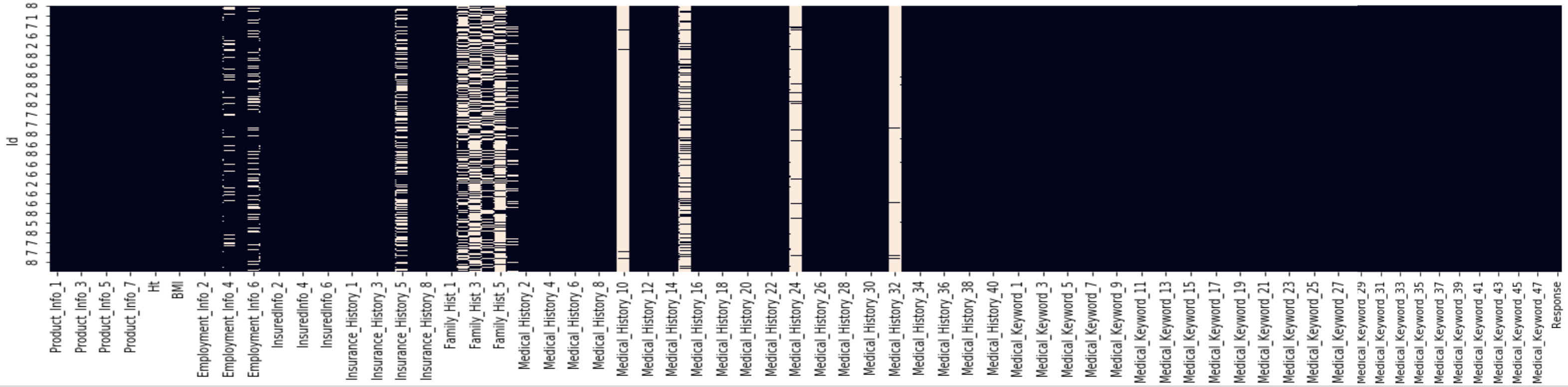
Data Type Split



Class Balance of Response Variable



Missing Value Report



Feature Variable	% Missing Values
Medical_History_10	99%
Medical_History_32	98%
Medical_History_24	94%
Medical_History_15	75%
Family_Hist_5	70%
Family_Hist_3	58%
Family_Hist_2	48%
Insurance_History_5	43%
Family_Hist_4	32%

Baseline model

Observations:

- Once we had the data prepared, we built a baseline models using 8 classification algorithms. This was done in order to have a benchmark model to compare with.
- However, the results were not satisfactory.

- We fitted three variants of Naïve Bayes classifiers, however, that also produced a similar result where the test scores were too low
  - BinaryRelevance
  - ClassifierChain
  - LabelPowerset

Result of Baseline Models

	Model_Name	Precision	Recall	Train_Accuracy	Test_Accuracy	F1_Score
1	RandomForestClassifier	0.47	0.47	0.8	0.47	0.47
2	GradientBoostingClassifier	0.47	0.47	0.48	0.47	0.47
3	XGBClassifier	0.47	0.47	0.57	0.47	0.47
4	AdaBoostClassifier	0.46	0.46	0.46	0.46	0.46
5	BaggingClassifier	0.43	0.43	0.78	0.43	0.43
6	DecisionTreeClassifier	0.37	0.37	0.8	0.37	0.37
7	LogisticRegression	0.34	0.34	0.34	0.34	0.34
8	SVC	0.18	0.18	0.18	0.18	0.18

Result from Naïve Bayes Classifiers

```
BinaryRelevance(classifier=GaussianNB(priors=None, var_smoothing=1e-09),
                require_dense=[True, True])

0.5300389393658446

ClassifierChain(classifier=GaussianNB(priors=None, var_smoothing=1e-09),
                order=None, require_dense=[True, True])

0.5300389393658446

LabelPowerset(classifier=GaussianNB(priors=None, var_smoothing=1e-09),
              require_dense=[True, True])

0.4157240867791582
```

Understanding Data – Exploratory Analysis

REVA UNIVERSITY

Highlights

- The data we have used is a collection of policy level information of random customers.
- Each observation has been classified on three risk categories by the underwriters.
- The data broadly contain the following feature-sets: Body structure, Product, Employment Information, Insurance History, Family History and Medical History

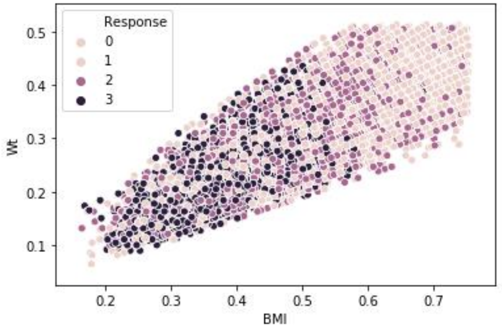
Descriptive Statistics

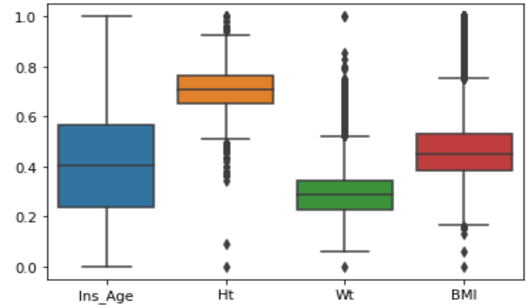
	Product_Info_4	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_4	Employment_Info_6	Medical_History_1
count	35953.000000	35953.000000	35953.000000	35953.000000	35953.000000	35953.000000	35953.000000	35953.000000	35953.000000
mean	0.342919	0.429306	0.708683	0.294140	0.470582	0.082957	0.007461	0.378181	8.021695
std	0.294550	0.194556	0.073842	0.088794	0.121112	0.088230	0.035481	0.354785	12.974841
min	0.000000	0.000000	0.000000	0.064854	0.151567	0.000000	0.000000	0.000000	0.000000
25%	0.076923	0.268657	0.654545	0.228033	0.388515	0.038000	0.000000	0.070000	2.000000
50%	0.230769	0.447761	0.709091	0.288703	0.454733	0.061800	0.000000	0.250000	4.000000
75%	0.487179	0.582090	0.763636	0.349372	0.533838	0.100000	0.000000	0.600000	10.000000
max	1.000000	1.000000	1.000000	0.828452	1.000000	1.000000	1.000000	1.000000	240.000000

Correlation Plot

	Product_Info_4	Ins_Age	Ht	Wt	BMI	Employment_Info_1	Employment_Info_4	Employment_Info_6	Medical_History_1
Product_Info_4	1.000000	-0.301644	0.125037	-0.044115	-0.138894	0.344285	0.039712	0.233310	0.058122
Ins_Age	-0.301644	1.000000	0.025706	0.123066	0.143475	0.073569	0.140789	0.364370	-0.107262
Ht	0.125037	0.025706	1.000000	0.617337	0.133398	0.194375	0.014956	0.098359	0.048498
Wt	-0.044115	0.123066	0.617337	1.000000	0.855395	0.091093	0.003314	0.016680	-0.021301
BMI	-0.138894	0.143475	0.133398	0.855395	1.000000	-0.009320	-0.006035	-0.043840	-0.057709
Employment_Info_1	0.344285	0.073569	0.194375	0.091093	-0.009320	1.000000	0.034297	0.373369	0.016479
Employment_Info_4	0.039712	0.140789	0.014956	0.003314	-0.006035	0.034297	1.000000	0.184324	-0.008093
Employment_Info_6	0.233310	0.364370	0.098359	0.016680	-0.043840	0.373369	0.184324	1.000000	-0.011645
Medical_History_1	0.058122	-0.107262	0.048498	-0.021301	-0.057709	0.016479	-0.008093	-0.011645	1.000000

Checking Spread of the Data





Data Treatment & Preparation

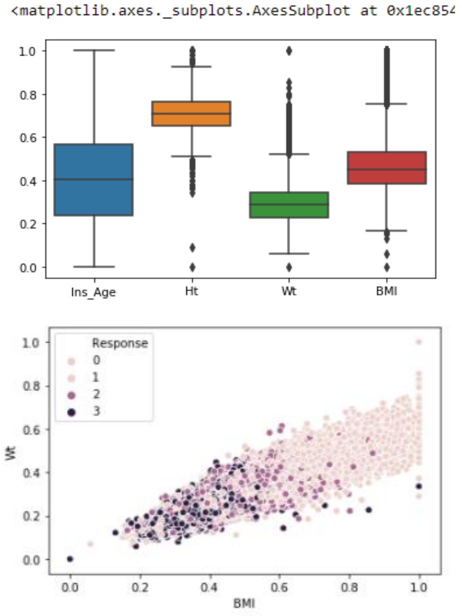
REVA UNIVERSITY

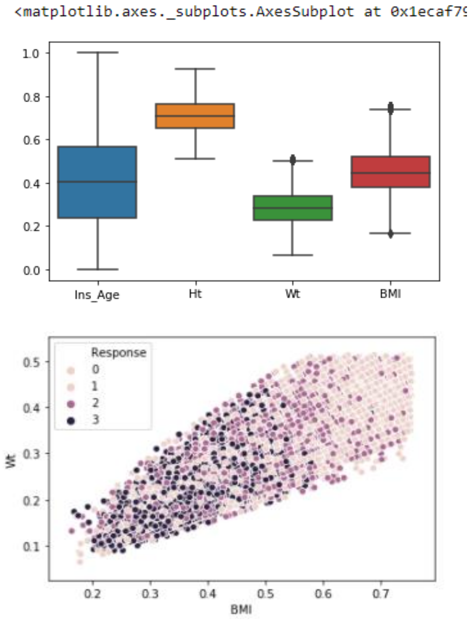
Imputing / Treating Missing Values

- Variables with more than 30% values missing were dropped.
- As remaining variables with missing values were categorical in nature and none of the imputation methods were proven to be useful, the respective rows were dropped.
- The methods were tried – KNN, Mode, SimpleImputer() from sklearn.impute

Treating Outliers

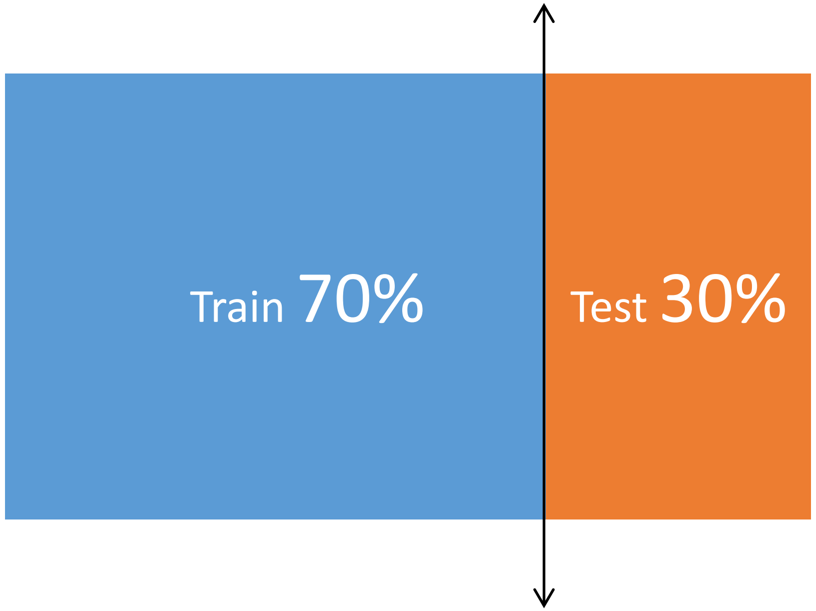
- Outliers were identified using by polling histogram and scatter plot
- The same was treated using the IQR flowing and capping methods, as shown below -





Splitting data into train and test

After reshuffling the observations the data was split into 80 – 20 for training and test purpose.



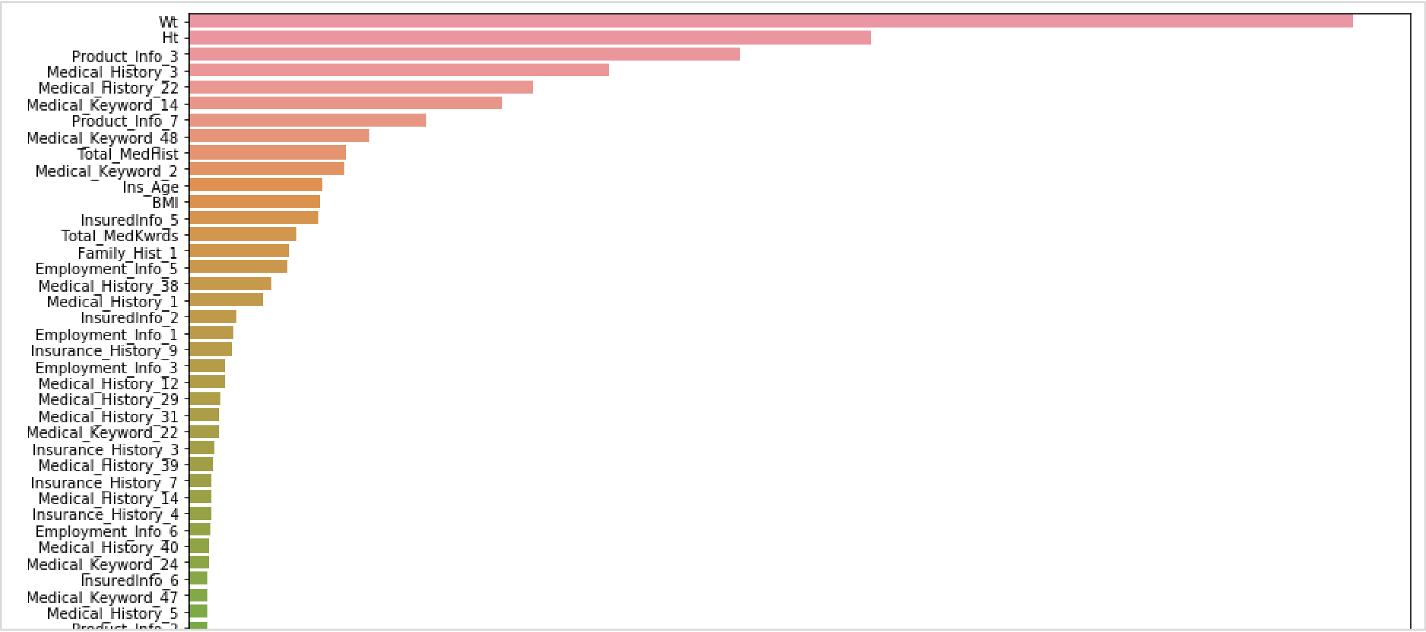
Feature selection

REVA UNIVERSITY

Notes / Observations:

- The next step was to identify the list of significant features. To do so, the following feature selection approaches were tried
  - Feature identification using Radom forest classifier
  - Chi -square test for independence
  - Select from model
  - Recursive Feature Elimination (RFE)
- Finally, after various iterations, we combined the results from two different approaches, i.e. RF and RFE

Features Identified by Random Forest



Features Identified by RFE

	Feature	Importance
1	BMI	11.690595
2	Wt	8.254409
3	Product_Info_4	5.556538
4	Ins_Age	5.087508
5	Employment_Info_1	4.270304
6	Medical_History_1	3.743639
7	Employment_Info_6	3.723107
8	Medical_History_2	3.649597
9	Ht	3.608519
10	Product_Info_2_en	3.103485

Model Interpretation and Conclusion

REVA UNIVERSITY

Notes / Observations:

- Once we had the final list of significant features ready, we re-ran all the models to compare the results
- Based on the outcome, we concluded that XGBClassifier is the best model in our case (predicting Risk Buckets), as it fitted the data well and showed promising result
- To test the selected model, we ran multiple iterations by -
  - reshuffling the samples
  - changing the proportion of training and test dataset
  - adding/removing variables that are not significant

  and after every iteration XGB classifier came out to be the best model with highest Accuracy and Precision
- In near future, we intend to feed in the fresh unseen data into the model to test its predictive capability

The Final Result

Model_Name	Precision	Recall	Train_Accuracy	Test_Accuracy	F1_Score
XGBClassifier	0.74	0.72	0.81	0.74	0.71
RandomForestClassifier	0.67	0.67	1	0.67	0.67
GradientBoostingClassifier	0.67	0.67	0.68	0.67	0.67
AdaBoostClassifier	0.66	0.66	0.65	0.66	0.66
BaggingClassifier	0.64	0.64	0.99	0.64	0.64
DecisionTreeClassifier	0.57	0.57	1	0.57	0.57
LogisticRegression	0.5	0.5	0.5	0.5	0.5
SVC	0.39	0.39	0.39	0.39	0.39

Test Accuracy = 0.74
Precession = 0.74
Recall = 0.72
F1 = 0.71

Future Scope

REVA UNIVERSITY

- ❑ Understand which categorical features have orders in them
- ❑ Using K-fold validations with at least 10% of the data in validation set
- ❑ Using Deep Learning Models
- ❑ Precision and Recall trade-off

Thank You!

Q&A

Code Repo and Citations

REVA UNIVERSITY

GitHub link

https://github.com/AshokShetty/REVA/blob/master/CapstoneProjectLnHPolicyUnderwriting.ipynb

References

- Accenture. 2015. “Harnessing the Data Exhaust Stream: Changing the Way the Insurance Game Is Played.” *Accenture Publication*
- Aggour, Kareem S., Piero P. Bonissone, William E. Cheetham, and Richard P. Messmer. 2006. “Automating the Underwriting of Insurance Applications.” AI Magazine
- Balasubramanian, Ramnath, Ari Libarikian, and Doug McElhaney. 2018. “Insurance 2030 – The Impact of AI on the Future of Insurance.” Digital McKinsey & Company
- Biddle, Rhys, Shaowu Liu, Peter Tilocca, and Guandong Xu. 2018. “Automated Underwriting in Life Insurance: Predictions and Optimisation.” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics).