A Project Report on

Harnessing the power of Data Science in policy underwriting

Submitted in partial fulfilment for award of degree of

MBA

In **Business Analytics**

Submitted by

Ashok Shetty

R17DM016

Under the Guidance of

Akshay Kulkarni

Lead Data Scientist

40 under 40 Data Scientist | Speaker | Author | AI Guest

REVA Academy for Corporate Excellence

REVA University

Rukmini Knowledge Park, Kattigenahalli,

Yelahanka, Bangalore – 560064

September, 2020

REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, Ashok Shetty hereby declare that I have completed the project work towards the MBA in Business Analytics at, REVA University on the topic entitled 'Harnessing the power of data science in policy underwriting' under the supervision of Akshay Kulkarni (Faculty & Industry Advisor). This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year 2020.

Place: Bengaluru

Name of the Student:
Ashok Shetty

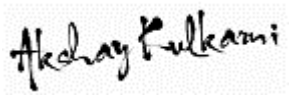
Date:

Signature of Student

REVA UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the PROJECT work entitled 'Harnessing the power of data science in policy underwriting' carried out by Ashok Shetty with SRN R17DM016, is a bonafide student of REVA University, is submitting the project report in fulfilment for the award of MAB in Business Analytics during the academic year 2020. The Project report has been tested for plagiarism and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of PROJECT work prescribed for the said Degree.



Akshay Kulkarni
Guide

<Signature of the Director>
<Name of the Director>
Director

External Viva
Names of the Examiners

1. Ravi Shukla – Sr. Advisor and Data Scientist, Dell
2. Krishna Kumar Tiwari, Senior Data Scientist, CoE, AI/ML, Jio

Place: Bengaluru
Date:

REVA UNIVERSITY
Bengaluru, India

Acknowledgement

I would like to thank our Chancellor Dr. P. Shyama Raju, Dr. S.Y. Kulkarni, Ex-Vice Chancellor Dr. D.K. Mallikharjuna Babu, Vice Chancellor and Dr. Dhananjaya, Registrar, for supporting the RACE program specifically designed for working professionals and providing facilities and infrastructure required and conductive conditions to offer the best learning experience. I am very happy to be called as a part of this program and REVA university.

Place: Bengaluru

Date:

REVA UNIVERSITY
Bengaluru, India

Similarity Index Report

Title of the Thesis: Harnessing the power of Data Science in policy underwriting

Total No. of Pages: 26

Name of the Student: Ashok Shetty

Name of the Guide(s): Akshay Kulkarni - Lead Data Scientist | Speaker | Author

This is to certify that the above thesis was scanned for similarity detection. Process and outcome are given below.

Software Used: Turnitin
Date of Report Generation: 06-Oct-2020 10:31PM (UTC+0530)
Similarity Index in %: 6%
Total word count: 4720

Place: Bengaluru
Date:

Name of the Student:
Ashok Shetty

Verified By:

Signature

Dr. Shinu Abhi, Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	AI	Artificial Intelligence
2	RPA	Robotic process automation
3	NBFC	Non-banking Financial Companies
4	SVC	Support Vector Classifier
5	LI	Life Insurance
6	RFE	Recursive Feature Elimination
7	IQR	Interquartile Range
8	XGB	XGBoost

List of Figures

List of Tables

Abstract

The ongoing digital transformation has hardly left any industry untouched. Also, the paradigm shift triggered by the pandemic, where most of the businesses are looking to shift from traditional platforms to digital platforms. In this journey of digitalization, insurance industry, which is generally known for its conventional way of doing business is also gearing up to jump into the race.

In a world with customized digital services and ever-changing buying behavior of customers the traditional ways of selling of L&H insurance is getting outmoded. All the possible info is collected from customers to analyze and understand underlying risks, which includes going through health checkups. The whole process takes approximately a month for the underwriters to make the decision. This causes loss of interest from customer's end. Which has resulted in about 60% of families in the country like US not having individual life insurance. Where the world sees this as a problem of 'protection gap', we see it as an opportunity to increase the market share by reducing the 'application to policy purchase' time.

With the help of a predictive model, which is capable of accurately classifying the risk using an automated mathematical engine and scientific approach, we can greatly impact public perception of the industry, increase our sales/profit and make the society more resilient by reducing the protection gap.

Contents

Chapter 1: Introduction

When the world is talking about Data Science, Artificial Intelligence and Big Data, there are certain industries (such as Insurance, Banking, NBFCs etc.) that are still responding to this revolution at a solver rate. They have a fair reason to do so. Their traditional business models have been doing well until recent times. In today's world most of the consumers prefer using online platforms, such as websites and apps to purchase/subscribe to various products, which due to which traditional ways of doing business are gradually becoming less and less effective. Also, the ongoing pandemic has had a long-lasting effect on consumers buying behaviours. This has certainly raised an alarm for all the product/services selling companies to modernise their ways of doing business.

The insurer are never known to take risks when the return is uncertain. However, there are certain factors such as socio-economic, regulatory, governmental, consumer behaviour and now the pandemic, that have turned many Life and Health insurance companies to technology to speed up the process and underwriting process cost effective (Bart 2012). Lately, the need for an automated underwriting system has gained some traction as insurers are looking to the reduction of workforce by reducing the "data to insight time" while upholding the accuracy of underwriting judgements.

One more reason for insurers to rely on digitization is to maintain profitability as they encounter rigorous socio-economic norms, while dealing with intricate books in a constantly challenging economic landscape. They understand the significance of establishing relations with policy holders to stay ahead in the competition. The insurance industry is now willing to leverage new developments in technical space, which includes AI, Big Data and Robotics to transmute policy selling framework and streamline overall operations (Albrecher et al. 2019).

The proposed solution in this article aims at automating a part of policy underwriting process of Life and Health insurance segment of an insurance company (Aggour et al. 2006). The proposed model will be used as centre piece in a broader automated solution, from the point of receiving an online application from the customer to selling him/her a personalized Life and Health insurance policy(Accenture 2015).

Chapter 2: Literature Review

In the past few decades Insurance companies have shown their resilience by withstanding various challenges faced by them. Still, they have started feeling the heat of competition from the tech savvy companies who are coming up with customized solutions. (Sharma 2019) (Cortis et al. 2019). Although transformation of core system has helped insurers reduce the gap, they are recognizing the competitive advantage which can be harnessed by adopting digital solutions (Nangla 2018) (Stoeckli, Dremel, and Uebernickel 2018). On the other hand, the consumers are showing interest in customized solutions. Also, the variety of risks being insured are ever-changing and offer fresh business prospects to insurance companies.

Within insurance, repetitive procedures like policy underwriting, claims management, accounting etc. still remain very tedious and mundane (Riikkinen et al. 2018). The traditional methods demand different systems (built on different technologies) like web, mainframes, programming languages etc. (Dubey et al. 2018). To automate the manual processes, its been observed that the insurers are now turning towards Machine Learning and Big Data systems. Insures who are adopting RPA are successfully able to accelerate their sales and are able to efficiently select the potential target segments. (Boodhun and Jayabalan 2018).

Insurance industry is currently at the brink of a complete digital transformation (Chae et al. 2001). Traditional business models are giving way to paradigms like,

- connected insurance
- usage-based premiums
- digital underwriting etc.

McKinsey's Global Institute report, 2017, Page 7, says that there is approximately fifty percent of scope for automation. In the next 5 years, up to 30 percent of the overall manual process in insurance industry would be automated or replaced with the help of RPA-driven tools (Hall 2017). The insurers are therefore persistently exploring processes for smart automation as Robotic Process Automation (RPA) along with Artificial Intelligence, Machine Learning and Cognitive tools that can be merged to gain productivity by reducing cost. Which makes the

future of underwriting very clear (Biddle et al. 2018), (Bonissone, Subbu, and Aggour 2002). (Balasubramanian, Libarikian, and McElhaney 2018).

Chapter 3: Problem Statement

In a world with customized digital services and ever-changing buying behavior of customers the traditional ways of selling of L&H insurance is getting outmoded. All the possible info is collected from customers to analyze and understand underlying risks, which includes going through health checkups. The whole process takes approximately a month for the underwriters to make the decision. This causes loss of interest from customer's end. Which has resulted in about 60% of families in the country like US not having individual life insurance.

Consumer expectations are increasingly being conditioned by the best practices found on sites such as Amazon, PayPal and eBay. Compared with these experiences, the traditional insurance process presents insurers with several challenges. The ones we are trying to address here are mentioned below -

1. In the majority of sales, insurance is purchased infrequently. In some lines of business, such as life insurance, it may only be bought once (and used only once!).
2. This lack of contact limits the opportunity for a distributor or an insurance company to establish a significant relationship with a customer and personalize the buying experience.
3. Besides, for underwriting risks and issuing policies, there is a list of tasks that are manually executed and are very tedious in nature.

This study explores options of using various Data Science techniques, specifically the Supervised learning methods to build an optimum model that could understand the underlying risk patterns and accurately classify the applicants into appropriate risk buckets. As this would be an automated solution, it would help insurers to address the challenges.

Chapter 4: Business Understanding

The speed and accuracy of the underwriting process is core to an insurer's success. With Nintex, now you can ensure that the appropriate policy and price are secured effectively and precisely, and in a way which is easy for a potential consumer to follow. The ongoing digitization of various platforms has changed the buying behavior of customers. Activities such as hiring cabs and communicating has changed the way life and health insurers used to conduct business. In near future, digital tools such as advanced analytics, cloud-based platforms and big data will empower a variety of business platforms by accumulating and scrutinizing massive amount of info for better sales strategy and underwriting. Leading insurers with digital platforms will lead the market by digitizing the salesforces, connecting with customers and liaisons in live omni-channel infra and offering robotic assistance 24x7.

Advancements in online platforms has been introduced at a time where the industry is encountering structural difficulties. Overall market share has been on a decline for the 3 decades so has been the detrimental progress between 2005 and 2015. Market penetration of new products has been on a decline from 17,000,000 per year in the 80s to 10,000,000 today.

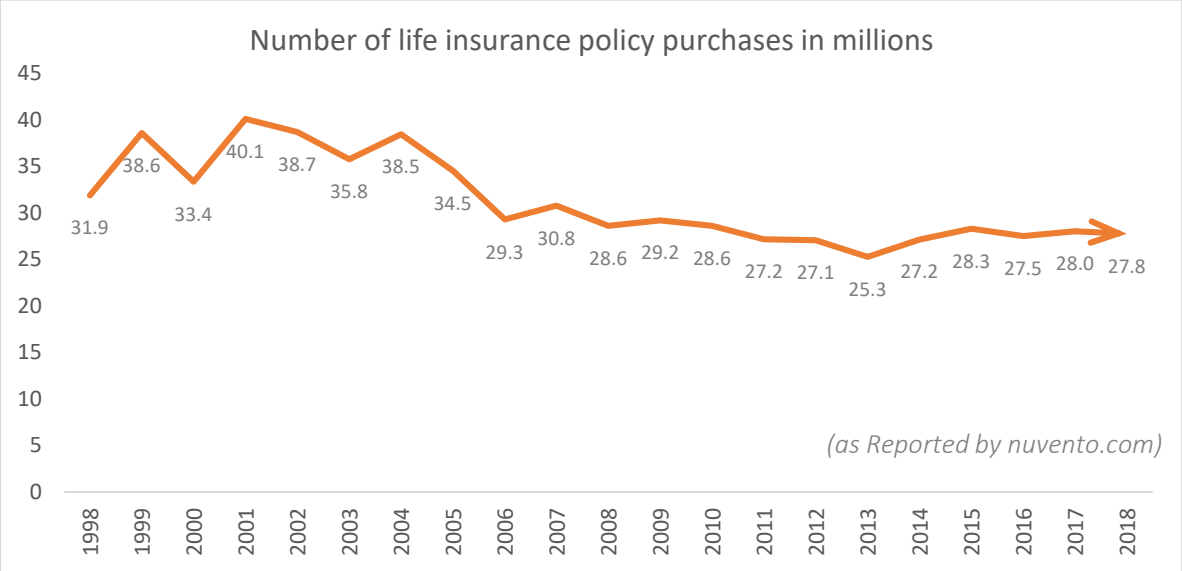


Figure 2 – YoY LI purchase volume

To make the best of these prospects, L&H insurers will have to outline an organized transformation strategy and digital approach. The multiyear plan enables leading insurers to create digital abilities in the correct sequence, enhance organizational formation to speed up invention, and implant the cultural and behavioral changes which are needed to grow sustainably.

With above motivation, a lot of life insurer are not planning to move their underwriting process to a digital platform, which could not only help them in increasing their market share by increasing velocity of sales, but also, let establish themselves as thought leader in the digitalization journey of insurance industry.

The above business problem can be answered with the data help of Data Science (Batty et al. 2010)., The public understanding of the industry can be significantly impacted by creating a predictive model which can precisely classify risks using a more scientific and programmed methodology. Which will increase our sales/profit and make the society more resilient by reducing the protection gap (Principles 2015).

The path to achieve the above is no different than other data science projects, where we started with the data gathering tasks till running multiple iterations of fine-tuned models to find the best fit (EIOPA 2019). The high-level project plan can be explained as below -

- Gaining in-depth understanding of the existence of business problem
- Estimating the potential commercial impact (saving cost / increasing revenue) of the proposed solution
- Gathering life insurance policy level data
- Anonymizing the data to get rid of personal and sensitive information
- Cleaning and treating the data to transform it into modeling-friendly format
- Trying various modelling approaches and interpreting the results
- Selecting and fitting the best suited model
- Running the results through industry/business experts to gather feedback
- Implementing changes suggested by the experts (if any)

Chapter 5: Objectives of the Study

The high-level objective of this project is to digitize the traditional policy underwriting process. with the help of existing analytical tools and techniques and reduce the response time, with an increase in the acquisition and retention rate. With this project we are aiming at offloading ~12% applications volume from underwriters' desk to an automated platform. The automated underwriting holds the promise of higher customer service levels, better process flow, improved tracking, enhanced agency communications, more consistent underwriting decisions and faster throughput — all resulting to the opportunity for substantially higher profits (Sigma 2015).

The above-mentioned high-level objective can be broken into smaller goals.

1. To increase 'From application to Quotation' time

This digitized approach aims at increasing the turnaround time of underwriting decisions. As Machine learning models are normally achieve swifter risk evaluations via a variety of algorithms and rule-based models, they would move applications through the system at a faster pace.

2. To come up with a cost-efficient way

The increased operational speed achieved from the proposed model would reduce the expenses usually related with policy underwriting. The cost saving would occur through automated and/or eliminated potential labor-intensive tasks. Furthermore, this tool would also reduce expenses through a more economically collected data, thereby, reducing tedious and costly measures taken by human underwriters to acquire the knowledge which is crucial for underwriting judgments.

3. To eliminate manual error and biases

An automated analyzation of a larger data would probably enhance the precision of underwriting capabilities. Moreover, the endless torrent of data collection would possibly detect unfavorable events prior to their occurrence. This model would also help in eliminating biases that are usually associated with manual policy underwriting. Elimination of the said bias would ensue via underwriting solutions made by following ML and rule-based algorithms instead of prejudiced assessments of a variety of risk aspects.

4. To make the decision making more agile

With the fact that the digital platform is capable of endless information compilation, which makes it feasible for insurers to revise contracts and fine tune the models as new risk factors appear.

5. To target digital platform users

Another objective of the proposed model is to attract new and young customers. Modernized & targeted policies are likely to attract millennials & GenZ segment.

6. To underlying trends and patterns of emerging risks

This tool could also possibly reduce the exposure of financial risk by detecting emerging risks before they get severe. The same can be done by constantly monitoring the patterns and trend in insureds' health status, personal behavioral data.

Chapter 6: Project Methodology

We used CRISP-DM methodology to execute this project structurally, where all the project activities were grouped into the following five predefined steps –

- Data understanding
- Data preparation
- Modeling
- Evaluation
- Development

Figure 1 – CRISP-DM Method

Each step mentioned above has been explained in detail in the following section.

A comprehensive information on the problem statement and business case has been gathered by interviewing team leaders and colleagues from the Life and Health department. After several meetings and brainstorming sessions, a desired format of the data along with potential features were finalized. However, extracting those features were not as easy as it looked as it contained some personal and sensitive personal information.

While collecting the data, since it had contained applicants' health and other personal information, we had to spend quite some time cleaning it. We identified personal and sensitive information manually and erased and/or masked them with some unique identifier. The next step was to get rid of any confidential information, such as the field and product names, which could expose some critical company information to unauthorized sources. Therefore, most of the feature names have been changed. Also, the observations have been masked to prevent from leaking any critical information.

Another major challenge was observed when we were trying to fit various models. Since all the observations in the data was purely generated through various underwriters, based on their individual experience and risk knowledge, identifying patterns between the response and explanatory variables proved to be a complicated task.

The complete data operations from the point of loading, cleaning, transforming enriching and modeling was done in Python.

Chapter 7: Data Understanding

The data described below represents various policy level information of the individual customers, who were classified manually by the underwriters based on various inputs received form the customers through application forms. This dataset contains about a hundred features showcasing applicant's demographic, family and medical history. The task is to come up with a classification method which can identify the right class of "Response" variable for the observations in the test set. "Response" is an ordinal measure of risk that has 3 levels.

Data dimension

Let's start with the shape of the data. It has 127 features (including response) with 59,981 observations.

Shape of Dataset (59381, 127)

Id	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_
8	1	D3	10	0.076923	2	1	1	0.641791	0.581818	0.148536	...	
4	1	A1	26	0.076923	2	3	1	0.059701	0.600000	0.131799	...	
8	1	E1	26	0.076923	2	3	1	0.029851	0.745455	0.288703	...	
8	1	D4	10	0.487179	2	3	1	0.164179	0.672727	0.205021	...	
8	1	D2	26	0.230769	2	3	1	0.417910	0.654545	0.234310	...	

Table 1 – Screenshot of raw data

Out of 126 exploratory variables, 18 features are numeric, 107 features are integer and1 feature is in object format. We can broadly categorize these variables into the following groups:

- Body structure
- Product
- Employment information
- Insurance History
- Family History
- Medical History

As part of the data masking process, to any linkage between individuals and their sensitive health related information, all the variables have been systematically masked. Numerical variables such as Height, Weight, BMI etc. have been normalized, whereas, all the classes in the categorical variables have been converted into integers.

Class balance

As shown below, the classes in the response variables were nicely balanced and hence the possibility of having problem of class-imbalance was eliminated. Proportionality balanced

classes were provided by the data pool team, who as per our request could pull reasonable number of samples for each class from their database.

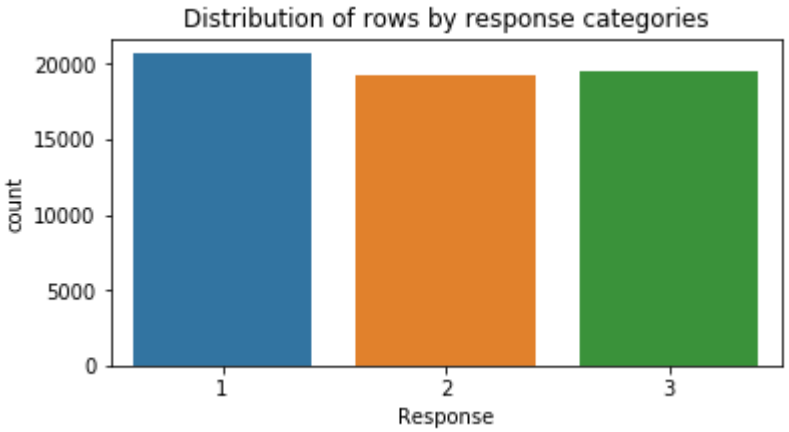


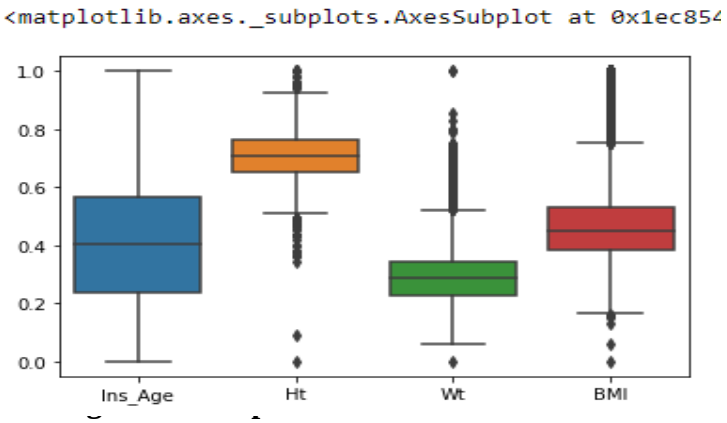
Figure 3 - Table 1 – Screenshot of raw data

Running descriptive statistics

Let us take a quick look at the measures of central tendency and dispersion. Since the data was normalized, it was a little hard to assess their shape and spread by looking at their means and standard deviation. And therefore, we started plotting them using Histograms.

	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

Table 2 - Descriptive statistics



The histograms revealed the problem of outliers. As you can see Height, Weight and BMI have a few observations that are spreading out of the interquartile range and hence need to be treated.

Chapter 8: Data Preparation

As we continued with the exploratory analysis, we started encountering certain issues in data such as missing values, multi-collinearity, outliers etc. This section talks about various treatments carried out on the raw data to deal with above mentioned challenges.

Identifying and imputing Missing values

There were about 13 features that had missing values. Out of which 9 had variables had more than 30% values missing. Due to the high percentage of missing values, all the 9 variables were dropped. To impute the missing values in the remaining features we used different imputation such as Regression and KNN.

	colsWithMissingValues	missingValuePercentage
0	BMI	0.031997
1	Employment_Info_3	11.416110
2	Employment_Info_5	18.278574
3	Insurance_History_4	42.767889
4	Family_Hist_1	48.257860
5	Family_Hist_2	57.663226
6	Family_Hist_3	32.306630
7	Family_Hist_4	70.411411
8	Family_Hist_5	14.969435
9	Medical_History_9	99.061990
10	Medical_History_14	75.101463
11	Medical_History_23	93.598963
12	Medical_History_31	98.135767

```
<matplotlib.axes._subplots.AxesSubplot at 0x1630f4baa48>
(35953, 118)
```

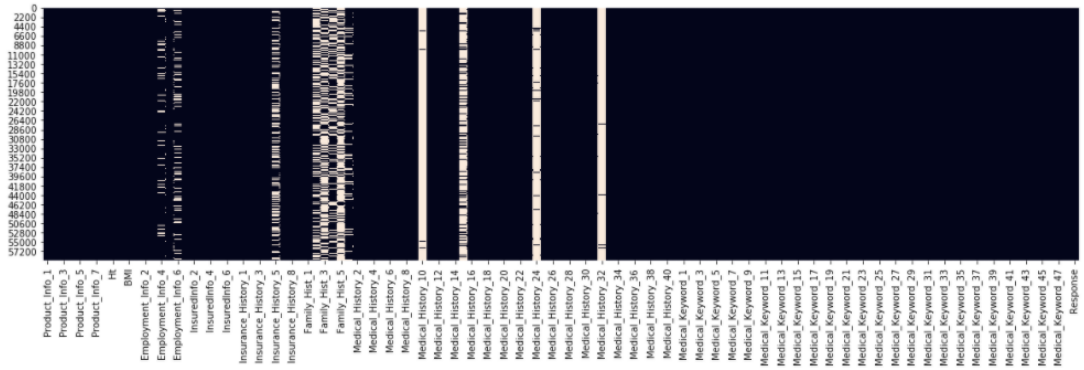


Figure 6 - Missing value report

Outliers

As mentioned in the previous section, when the numeric variables plotted, we noticed that there were a good number of observations outside the interquartile range (IQR). The same was treated using flooring and capping technique.

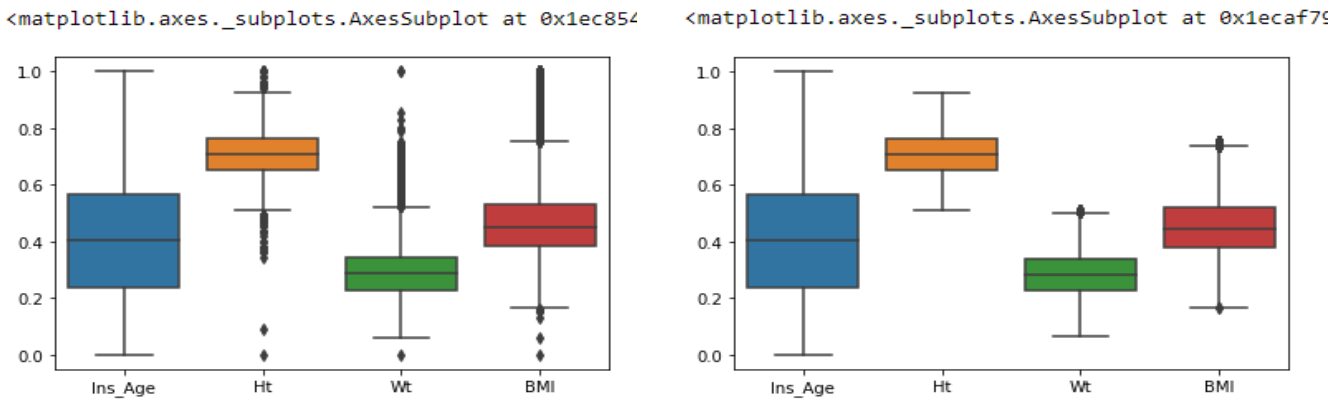


Figure 7 - Pre and Post the treatment

Checking for Correlation

Using Correlation and Scatter plot we were able to identify features that have linear relationship. As shown below, BMI and Weight have a strong positive correlation, whereas, there is no correlation between Weight and Height.

	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

Table 3 – Correlation Matrix

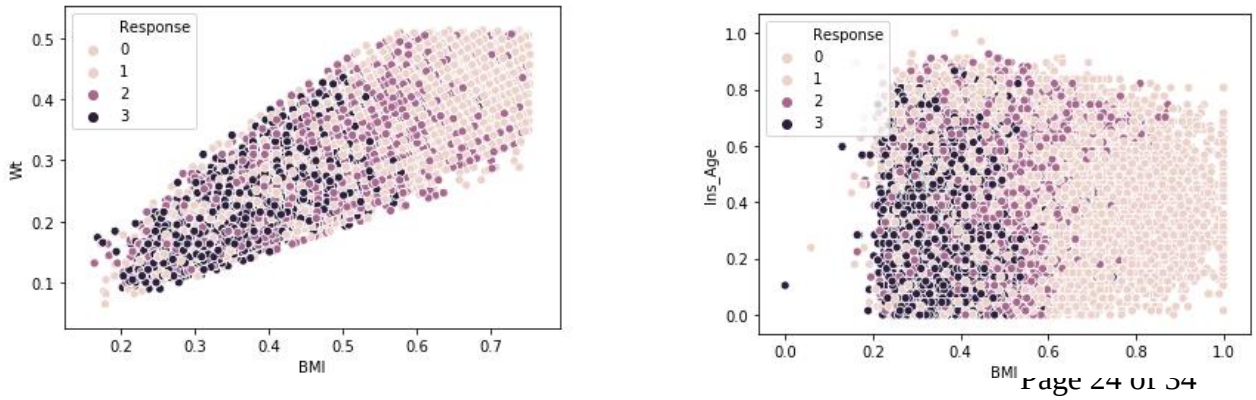


Figure 8 – Correlation plot

Feature selection

Selecting a set of significant features is an essential step before building any model, as it helps modelers in understanding the predictive power of each variables in question. As we have a good mix of both numeric and categorical variables in our data, we had to use different techniques to come up with our final list of variables. We used the following methods -

- Random forest (for both categorical and numeric variable)
- Recursive Feature Elimination (for both categorical and numeric variable)
- Select from model (for both categorical and numeric variable)
- Chi-Square (for categorical variables only)

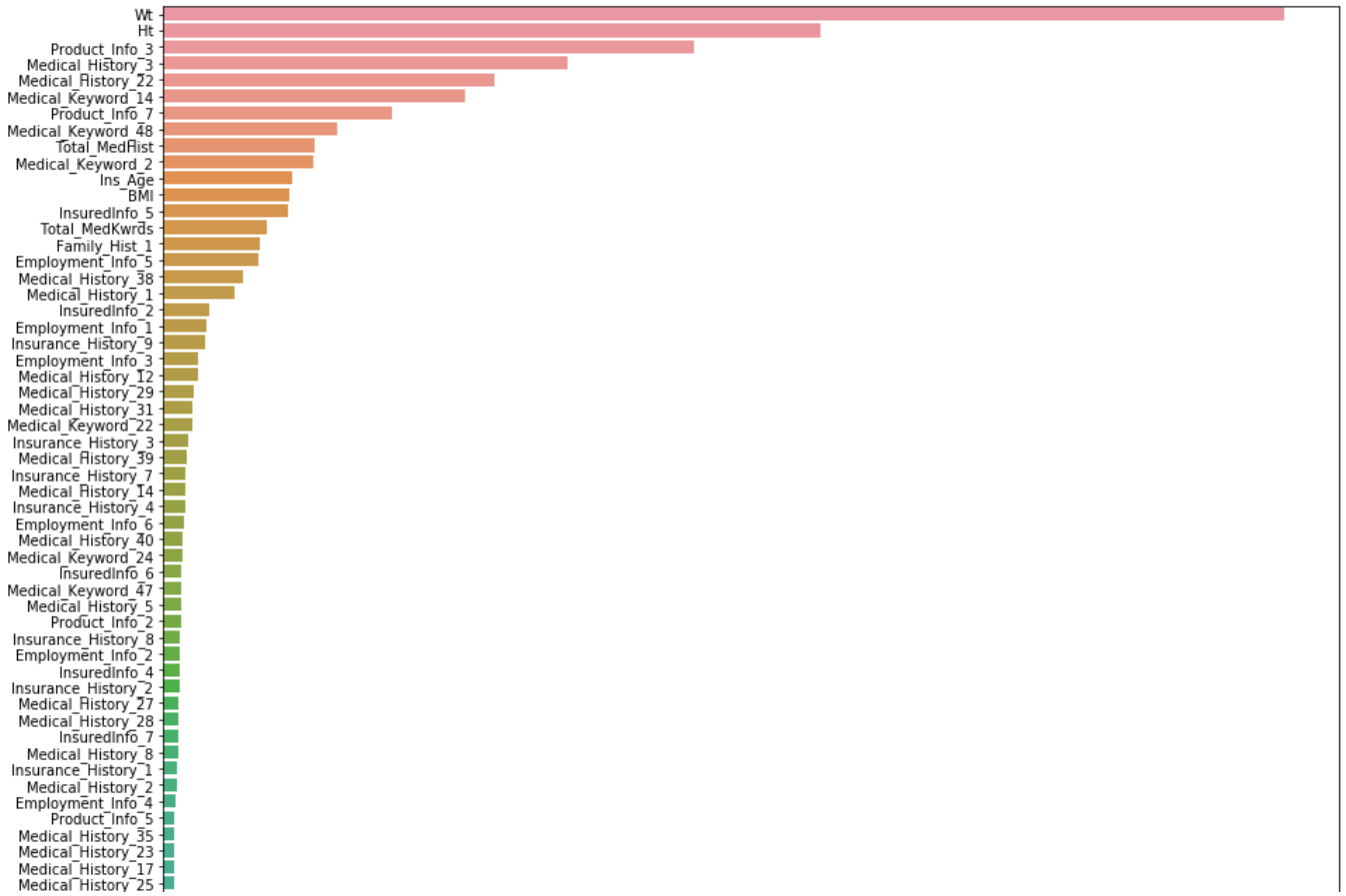


Figure 9 – Feature Importance

Chapter 9: Modeling

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. Here's the result –

	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

Table 4 – Baseline model result

The next step was to identify the list of significant features and the best algorithm/model that would predict the variation in the response variable well. 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)

Also, three additional models were built using the below approaches. This was done to check if that would produce any better result. However, as shown below the results were not satisfactory:

- BinaryRelevance
- ClassifierChain
- LabelPowerset

However, that also produced a similar result where the test scores were too low.

```
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
```

Figure 10 - Naïve Bayes model results

Finally, using the combination of Recursive Feature elimination and Random Forest, we were able shortlist the significant variables. Total 27 features were selected in building the final model. Here is the list of top 10 important features:

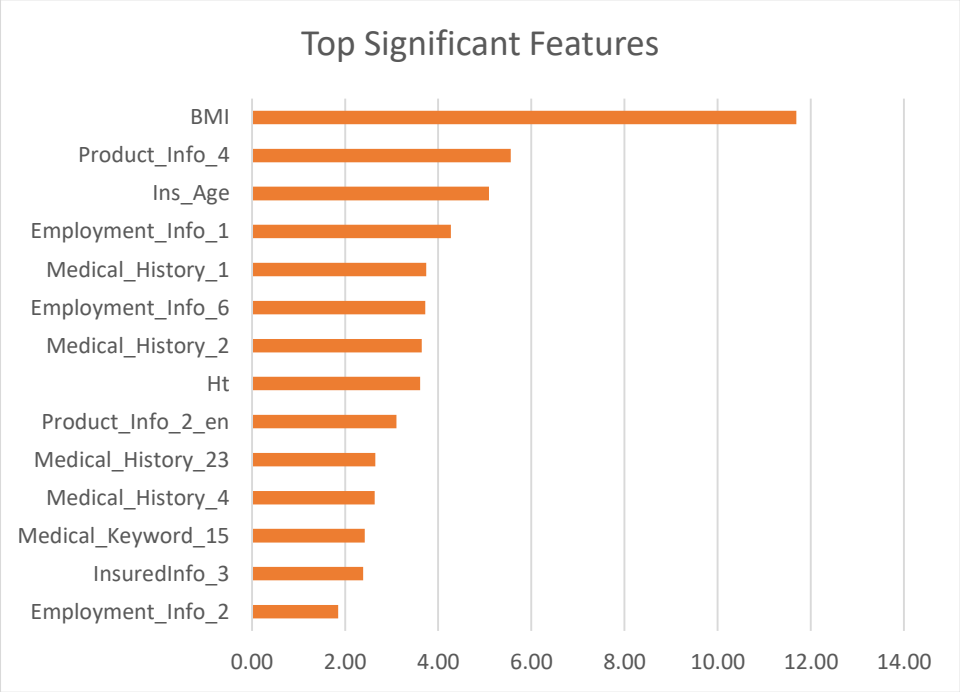


Figure 11 – Top Significant variables

Once we had the final list of significant features ready, we reran all the models to compare the results.

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

Table 5 - Result from final set of models

Based on the outcomes, we concluded that XGBClassifier is the best model in our case to make the policy related prediction, as it fitted the data well and showed promising results with,

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

Chapter 10: Model Evaluation

After identifying the best suited algorithm, now we had to test it against various hypothetical scenarios and therefore along with XGB, we re-ran all the other models by making following changes in the data -

- 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. XGB is an augmented boosting algorithm tailored to be dynamic as well as compact. This modelling technique employs Gradient Boosting module. XGB offers an analogous tree based (GBDT, GBM) that answers many machine learning challenges problems in a precise way.

In near future, we intend to feed in the fresh unseen data into the model to test its predictive capability.

Chapter 11: Deployment

As soon as we run few more checks on the model by feeding in fresh observation and re-evaluating the importance of selected features, the same will be shared with the underwriters to get their opinions. And once we have the go ahead, this model will be used as a center piece of an automated platform, which would receive online applications from various sources and based on the inputs will classify the customers into appropriate risk buckets.

Chapter 12: Conclusions and Recommendations for future work

Based on the results, we can conclude that out of all the features, height, weight, age, BMI along with some other key medical history have high influence on underwriting decisions. After trying almost all the classification technics and studying the results in detail, we can say that XGB classifier has learnt the underlying patterns well to make the prediction. It would however be interesting to see the model's performance on completely fresh data, as underwriters' decisions includes a lot of bias.

Just imagine a scenario, where you are feeding to the model with data captured from various sources and each source has its own biases. Similarly, in our case we were using the data collected from several underwriters, who using business knowledge and their own experience underwrite the cases. In such scenarios, it is very difficult to obtain high accuracy. However, in near future, we are expecting to receive the description of each feature, which was masked due to data protection issues. Once we have the description with us, it would be easy for us to process the training data. E.g. currently we are unable to differentiate between ordinal and nominal variables, which is an essential piece of information.

Bibliography

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*.

Albrecher, Hansjörg, Antoine Bommier, Damir Filipović, Pablo Koch-Medina, Stéphane Loisel, and Hato Schmeiser. 2019. “Insurance: Models, Digitalization, and Data Science.” *European Actuarial Journal*.

Balasubramanian, Ramnath, Ari Libarikian, and Doug McElhaney. 2018. “Insurance 2030 – The Impact of AI on the Future of Insurance.” *Digital McKinsey & Company*.

Bart, Patrick. 2012. “Insurance: Improving Claims Management.” *Http://Www.Theactuary.Com*.

Batty, Mike, Cera Arun Tripathi, Alice Kroll, Cheng-sheng Peter Wu, David Moore, Chris Stehno, Lucas Lau, Jim Guszcza, and Mitch Katcher. 2010. “Predictive Modeling for Life Insurance.” *Ways Life Insurers Can Participate in the Business Analytics Revolution*.

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)*.

Bonissone, Piero P., Raj Subbu, and Kareem S. Aggour. 2002. “Evolutionary Optimization of Fuzzy Decision Systems for Automated Insurance Underwriting.” in *IEEE International Conference on Fuzzy Systems*.

Boodhun, Noorhannah, and Manoj Jayabalan. 2018. “Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms.” *Complex & Intelligent Systems*.

Chae, Young Moon, Seung Hee Ho, Kyoung Won Cho, Dong Ha Lee, and Sun Ha Ji. 2001. “Data Mining Approach to Policy Analysis in a Health Insurance Domain.” *International Journal of Medical Informatics*.

Cortis, Dominic, Jeremy Debattista, Johann Debono, and Mark Farrell. 2019. “InsurTech.” in *Disrupting Finance*.

Dubey, Aman, Tejisman Parida, Akshay Birajdar, Ajay Kumar Prajapati, and Sagar Rane. 2018. “Smart Underwriting System: An Intelligent Decision Support System for

Insurance Approval Risk Assessment." in *2018 3rd International Conference for Convergence in Technology, I2CT 2018*.

EIOPA. 2019. "Big Data Analytics in Motor and Health Insurance: A Thematic Review." *EIOPA Thematic Review*.

Hall, Shanique. 2017. "How Artificial Intelligence Is Changing the Insurance Industry." *The Center for Insurance Policy & Research*.

Nangla, Karan. 2018. "Artificial Intelligence and Health Insurance." *Journal of the Insurance Institute of India*.

Principles, Climatewise. 2015. "Closing the Protection Gap." *Best's Review*.

Riikkinen, Mikko, Hannu Saarijärvi, Peter Sarlin, and Ilkka Lähteenmäki. 2018. "Using Artificial Intelligence to Create Value in Insurance." *International Journal of Bank Marketing*.

Sharma, Shibyanshu. 2019. "Artificial Intelligence in Insurance Sector." *Journal of the Insurance Institute of India*.

Sigma. 2015. "Life Insurance in the Digital Age : Fundamental Transformation Ahead." *Swiss Re Sigma*.

Stoeckli, Emanuel, Christian Dremel, and Falk Uebernickel. 2018. "Exploring Characteristics and Transformational Capabilities of InsurTech Innovations to Understand Insurance Value Creation in a Digital World." *Electronic Markets*.

Appendix

Plagiarism Report[1]

Publications in a Journal/Conference Presented/White Paper[2]

Any Additional Details

[1] Turnitn report to be attached from the University.

[2] URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.