12th Webinar on Health Care Insurance

August 23, 2024

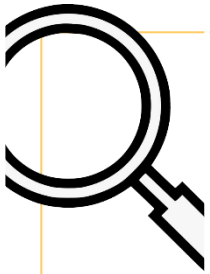
Renewal Propensity Modelling in Health Insurance

Sourish Chakravarti
Data Scientist, ManipalCigna Health Insurance

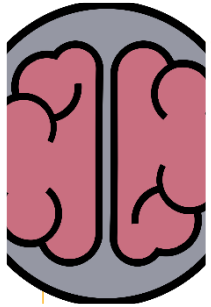
Institute of Actuaries of India

Problem Statement

Low customer persistency in health insurance poses a major challenge in building a sustainable portfolio in the long run



Identify segment of customers who are less/more likely to renew, prior to their renewal period



Understand the mindset of non-renewing & renewing customers to provide better service

Method 1

This can be done using basic statistical tools. E.g., one-way/ two-way

Method 2

However, there are multiple parameters, hence a multivariate analysis is better. E.g., GLM/ ML modelling

Brushing up some ML Concepts

Supervised v/s Un-supervised

- **Supervised learning** has labelled data. i.e. There are (X) inputs and (Y) output.
- Goal is to predict Y basis the X variables – learn the function Y = f(X)
- **Unsupervised learning** has un-labelled data. i.e. There is no output (Y) variable.
- Goal is to detect the underlying pattern and structure of the data

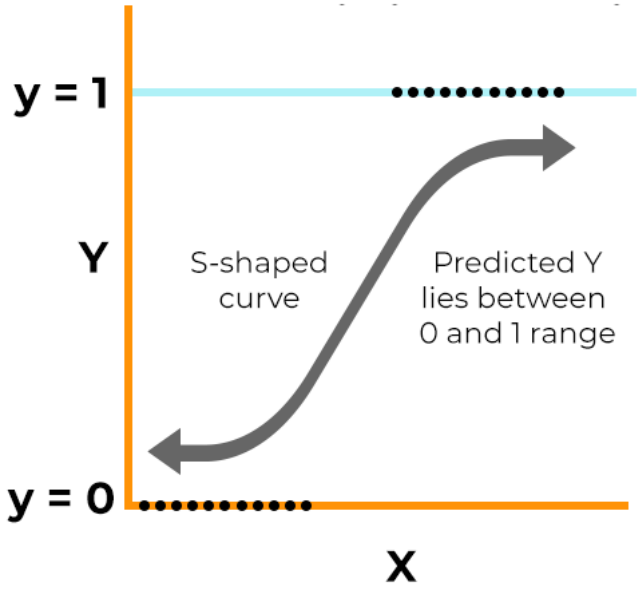
Classification v/s Regression

- **Regression algorithm** is used when our target variable (Y) is continuous in nature. E.g., Age, Salary etc.
- **Classification algorithm** is used when our target variable (Y) is categorical in nature. E.g., Gender, Fraud (Y/N), Renewed/Non-Renewed. Classification can be:
  - Binary: Classification task with two possible outcomes
  - Multinominal: Classification with more than two classes

Commonly used Algorithms

- Unsupervised:
  - K-means, K-mode, Hierarchical Clustering etc.
- Supervised:
  - GLM (logistic regression), SVM, KNN, Decision Tree – Random Forest, XG-Boost etc.

Classification using Logistic Regression



- Logistic Regression uses sigmoid function (S shaped curve) to convert the predicted output (Y) into a probability (between 0 and 1)

$$Y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

- Coefficients of logistic regression are **Log of Odd Ratios**

Few Data assumptions while using Logistic regression

Linearity: Predictor should be linear with the link-function of Y

No Perfect Multicollinearity: No X variables should be perfectly correlated to each other

Independence: Observations should be independent of each other. i.e., residuals should not be correlated

Advantages

- Simple and interpretable
- Efficient - Fast
- Can handle multiple classes

Disadvantages

- Requires large data set
- Outlier sensitive
- Cannot handle non-linear data

Data Creation and EDA

Step 1: Data Preparation:

- **Target variable** – Renewal Flag
- **Predictors:**
  - Policy details, Customer Demography, Distribution
  - Service-related information
  - Claim experience
  - External data – Credit Rating etc.

	Target (Y)	Predictors (X)		
Policy	**Renewal**	**Age**	**Gender**	**…**
a1	Y	35	M	..
a1	Y	17	M	..
b1	N	22	M	..
b2	Y	56	F	..

Main Dimensions to capture:
1) Affordability of customer 2) Willingness & Interest of customer

Data Cleaning and Data massaging:
Includes outlier detection and removal, missing value treatment

. **Step 2: EDA and Feature Engineering:**

Exploratory Data Analysis (EDA) is the process of analysis and visualizing the dataset to understand:

- Any underlying trends, seasonality
- Relationship and pattern among the predictor variables (X)
- Relationship and pattern among the target (Y) and predictor (X)
- Outlier and anomalies

Tools used for EDA include:

- Correlation matrix/ Correlation Plot
- One-way/ Two-way tables.
- Histogram & line charts

Feature Engineering and Segmentation

Step 2: Feature Engineering:

Feature Engineering: process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model performance on unseen data

Example:

1) Portability and Policy Year
2) Customer Touchpoints
3) Riders attached with the base policy

Segmentation

Data may show materially different renewal behavior for:

1) Different Products – E.g., HNI v/s Basic Product
2) Different Regions – E.g., Tier 1 v/s Tier 2 City
3) Different time periods. – Covid v/s Post Covid

Solution:

1) Separate Models
2) Segment as a flag

Proper EDA and Feature Engineering will ensure that data created is in line with business intuition and understanding.

Model Development

Step 3: Model Development:

Once the data is ready, following model steps are used:

- Divide dataset into **training data (70%)** and **test data (30%)** in a statistically random manner
- Running the model and analyzing the output –
  - Variable coefficients
  - Variable importance (p-value)
  - Model KPI : Confusion Matrix and its derivatives like accuracy, specificity, sensitivity etc.

.

- The model provides its output (Y-pred) in terms of a probability – **propensity score**

- This score is converted into a class label, governed by a parameter known as the decision **threshold** - 0.5 is the default for normalized predicted probabilities
- For our case if output >= 0.5 = Class 1 (Renew), output < 0.5 = Class 0 (Non-Renew)

*Confusion Matrix**		**Predicted**	
		Renewed	**Non-Renewed**
Actual	**Renewed**	**TP**	**FN**
	Non-Renewed	**FP**	**TN**

Vetting the actual and predicted output we get multiple **KPI** like:
Accuracy: (TP+TN) /(TP+TN+FP+FN)
Sensitivity: TP/(TP+FN)
Specificity: TN/(TN+FP)

***Note : TP : True Positive, TN : True Negative, FP : False Positive, FN : False Negative**

Inference & validation

Step 4: Model tweaking and Validation:

Default threshold (0.5) may not represent an optimal interpretation, due to:

- The class imbalance in data
- The cost of one type of misclassification is more than another type of misclassification

Optimal Cut-off can be selected using the following methods:

- **ROC (Receiver Operating Characteristics)**
  - ROC helps to find the optimal cut-off that maximizes TPR and minimizes FPR*
- **Decile / Percentile**
  - Ranking the propensity scores, we can get the top/ bottom x percentile of customers basis their propensity of renewal

Out of Sample Validation:

When the model is ready, the final step is to run the model on a out of sample data-set (usually 3-6 months) to ensure:

- Model output is stable over time
- Model output is in lines with expectations

Step 5: Model Maintenance:

Once deployed, model needs to be retuned, refreshed on a periodic basis to:

- Increase the available data points (More the data the better)
- Capture any recent renewal pattern and trends

***Note : TPR – True Positive Rate = TP / (TP + FN) | FPR – False Positive Rate = FP / (FP + TN)**

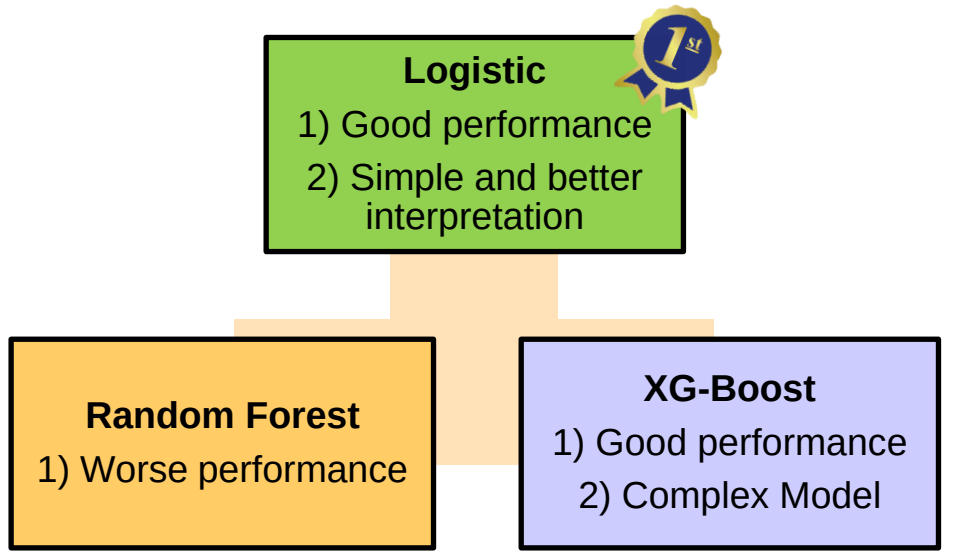
Choosing Best Model - Why Logistic?

During development of the renewal model, multiple algorithm needs to be tested. For our case, the following were tested:

1) **Logistic Regression**
2) **Tree Based Model**
   I. **Random Forest**
   II. **XG-Boosting Model**

Things to keep in mind while selecting models:

1. **Final Model Performance** – Accuracy, TPR, FPR, AUC-ROC etc.
2. **Target variable balance** – Tree models are better with class imbalance
3. **Overfitting, Underfitting**
4. **Model simplicity** – Ease of interpretation



Hence, Logistic Regression is chosen for our use case

Predicting Renewal – Action Items

Once the model is ready and deployed, it can be used (scored) on set of customers who are **due for renewal.**

Green customer

- Customers with high propensity score- Highly likely to renew

Amber customer

- Customer with average propensity score- Slightly less likely to renew.

Red customer

- Customer with low propensity score- Least likely to renew/ high lapse rate

Basis the optimal cut-off chosen, the output for the customers can be split into 2 or more sections to segment the customers according to their renewal behavior

Use Case:

1) **Increase renewal rate by designing retention strategy for the Red Customers**
2) **Optimize Operational Cost for each segments**

Key Take-away

Success of the model depends on:

- The variety of data available (data depth)
- The usability of the available data
- Robust model maintenance to capture new patterns

- ML Models are trained on existing renewal patterns
- Patterns may vary for separate cohorts. Hence data engineering and segmentation may be necessary

Predictive quality depends more on data than on algorithm

- There is no single BEST algorithm
- Selection of model varies case by case basis.

Model Output can be used to

- Improve Retention rate
- Improve operational efficiency