SOCIETY OF ACTUARIES®

SOA Predictive Analytics Seminar – Taiwan

31 Aug. 2018 | Taipei, Taiwan

Session 4

Case Study of Modern Approach to Lapse Rate Assumption

Richard Liao, ASA
Stanley Hsieh

Case Study of Modern Approach to Lapse Rate Assumption

RICHARD LIAO/ STANLEY HSIEH

31 August, 2018

SOCIETY OF ACTUARIES®

Table of Contents

Why machine learning for lapse study?

SOCIETY OF ACTUARIES®

What is Machine Learning?

- Use statistic to give computer ability to learn
- Let the algorithm do the job to improve the prediction

What is Machine Learning?

Supervised learning

- Learning a function with input and output
- Labeled training data set is used to learn a function
- This function can be used to map new examples

Unsupervised learning

- Learning a function describing the structure of unlabeled data

What is Machine Learning?

Regression

- To predict "continuous" outcomes

Classification

- To predict "discrete" classes

Training set

- For training machine learning model

Validation set

- For machine learning model adjustment

Testing set

- For prediction and testing prediction power

What Impacts Lapse Rate?

- What are the attributes affecting lapse rate?
- Only one attribute or more attributes?
- Should it be really time dependent?
- Different product types?
- Sales channel or even sales office, sales person?
- Social economic trends impact?
- Other factors we don't normally think of?

Traditional Experience Study

- Traditional way of lapse rate experience study usually contains a few dimensions only:

Premium mode	Policy year	Product type	Gender	Sales channel

- Often times, the result by the above dimensions look volatile. Should more dimensions be considered? What are those? How can we find them easier?

Business Impact by Lapse Rate

- It is really, really hard to sell an insurance policy. Have we tried upmost to prevent lapse?

Business Impact by Lapse Rate

Profit and Loss

- High volatility of lapse rate estimation may cause high volatility of profit and loss, especially after the implementation of IFRS17, significant difference of actual lapse realized and expected lapse becomes the source of profit and loss

Market influence

- The ability to monitor and retain insurance policies may influence the domination of market share and corporate reputation

Customer value

- When high value policies are sold, preventing policies from surrender is the key to keep customer value or company value

Business Impact by Lapse Rate

Marketing strategy

- When knowing the possible lapse behaviors resulting from specific product types, sales behaviors, policyholders' features, non-policyholders' features, or other factors, insurance companies can have better position on making marketing strategy for policy sales

Product design

- Lapse rate plays a key role when pricing a product and determining the profitability of a product. Accurate estimation of lapse rate becomes important when implementing business plan

Risk management and ALM

- Asset and liability management and risk capital management heavily relies on the accuracy of cash flow projection. Hence, lapse rate prediction is extraordinarily crucial for the management decision

Linking Machine Learning with Lapse Study

- Supervised learning X ⟶ Y
- Binary classification problem:

  Y = 1 for Surrender

  = 0 for Non-surrender
- Combine policy related data with economic data to enrich data
- Algorithm learns from information of data
- Select an appropriate machine learning model

Benefit of Machine Learning Approach

Higher prediction power	More dimensions to determine lapse behaviors
More automatic assumption making process	Improve short term money management

Machine learning preparation

Project Flow

Problem Definition → Data Investigation → Modeling → Analysis

Data, Resource and Business Impact

- Data availability
  - Cost of data purchase or collection
  - Privacy issue / legal issue
- Data quality
  - Consistency over time regarding definitions
  - Mindful of “garbage in, garbage out”
  - Enough data counts
  - Enough variable (attribute) counts
  - Dealing with missing date – apply common methodologies
- Investment in data infrastructure

How to succeed?

- Start from small and realistic goals, and build from the success to make it bigger
- Cooperate with subject matter experts
- Understand the implementation needs of the model, such as purpose, cost, time frame of each prediction, or resource supported

Data Types & Variable Types

- Independent Variable (X):
  - Policy Related Data:
    premium balance, channel mode…etc
  - Economic Index:
    GDP, stock index, inflation, real-estate price…etc
- Dependent Variable (Y):
  Y = 1 for Surrender and Y = 0 for Non-Surrender

Quality of Data & Data Collections

- Source of Data: Internet? Agent?
- Why do we have missing data?
- There is no value in learning constant data
- Some data is recorded recently so there is lack of historical data
- Communication with data engineer for data cleaning
- Actuarial Perspective is important for variables selection

Data Cleaning Techniques & Transformation

- Select a threshold for excluding variable with too many missing data
- Mean Imputation – by filling data mean to missing observations
- We can use feature engineering to create variables
- Categorical variable has to be transformed into factors

Machine learning model

SOCIETY OF ACTUARIES®

Machine learning – Model

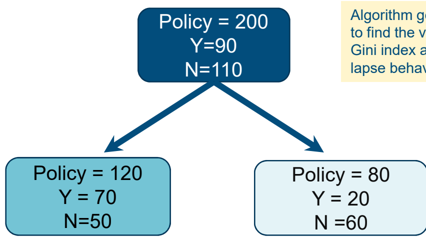
Generalized Linear Model	Decision Tree
Random Forest	Gradient Boosting Machine

Generalized Linear Model

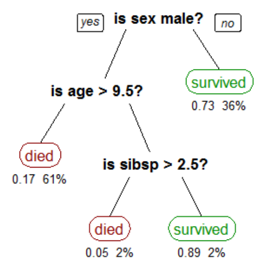
- Result can be interpreted by coefficients of variables
- Link Function and Distribution – logit and binomial for binary classification
- Classical Way – By using statistical test for model significance
- Machine Learning Way – By feeding more variables for prediction power
- Regularization: To control overfitting of GLM
- Regularization tool: Ridge (L2-norm) vs Lasso (L1-norm)
- LASSO is widely more popular due to its penalty character

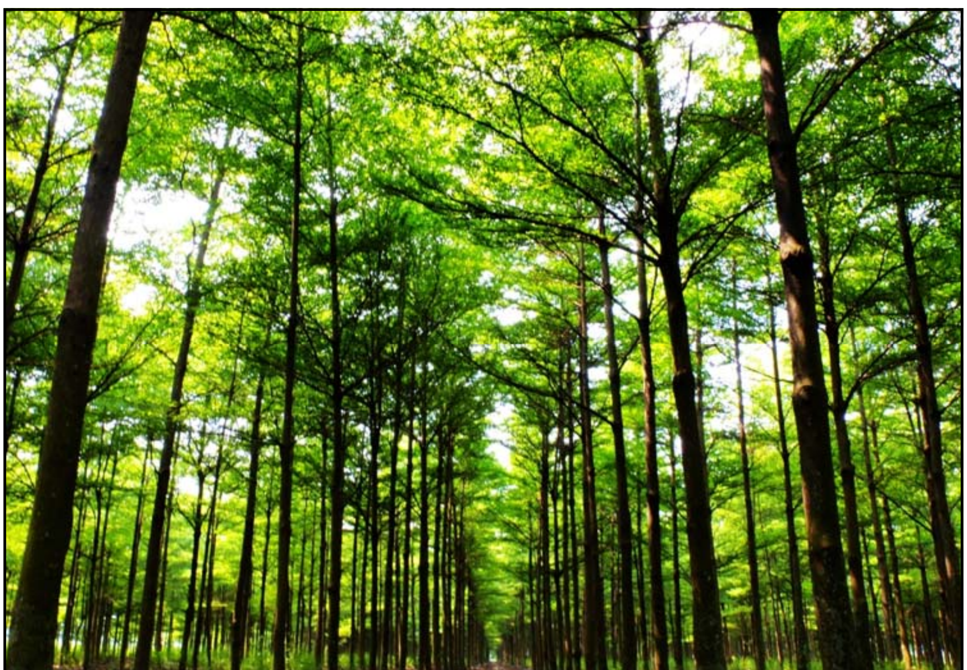
Decision Tree

- Decision boundary is drawn to capture non-linear trend
- Key idea of algorithm: recursive binary splitting
- Measure impurity of node by Gini Index



Algorithm goes through the variables to find the variable that has lower Gini index as this variable classifies lapse behavior more distinguishably.





Random Forest

- Start from idea of bagging – resampling and bootstrapping
- Searches for the best feature among a random subset of features – to de-correlate the trees
- Trees can be implemented by parallel computation

Gradient Boosting Machine (GBM)

- G(x) = F(X) + h(x) + ……
- F(X) = weaker learner
- Residuals = y – F(X)
- Residuals is trained in the direction of gradient descent
- Add the trained residuals to weaker learner then repeat this process
- Train a "bad" tree first then train its residual to make it a better tree
- Generally, a powerful machine learning model

Case study – analysis of outcome

Outcome

- Class Probability:

  p0 = Non-surrender probability and p1 = Surrender Probability
- Optimal Threshold – Threshold that optimally decide whether each policy will surrender next quarter

Predict	p0	p1
0	0.99	0.01
0	0.90	0.10
1	0.11	0.89
0	0.91	0.09
0	0.87	0.13
0	0.88	0.12
1	0.12	0.92

Metrics

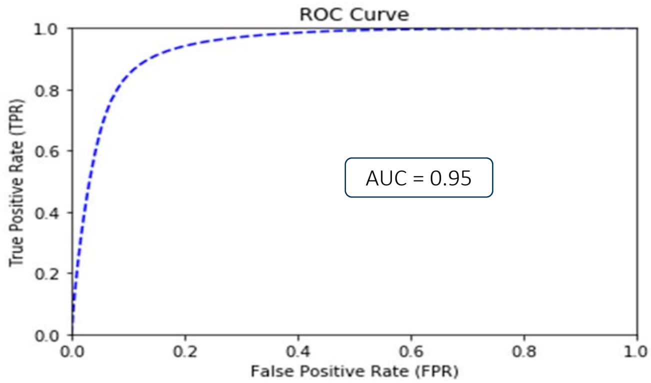
- To evaluate performance of model
- To prevent overfitting
- MSE (Mean Square Error):

It can be used to evaluate numeric prediction like stock price prediction

- AUC (Area under Curve):

This is what we used for the case study which is a classification problem.

AUC (Area under Curve)

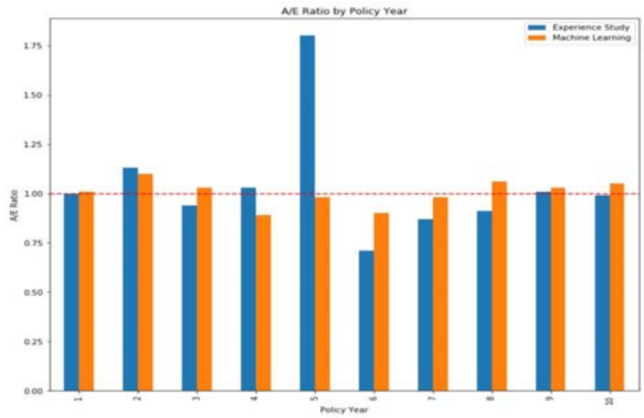


- AUC stands for Area under the ROC (Return of Characteristics) Curve
- Points on ROC is the False Positive Rate and True Positive Rate at certain threshold

Hyper-Parameter Tuning

- Maximum Variables Allows in a GLM :
  Tradeoff between model explanation and model prediction
- Depth of Tree:
  Is deeper the tree better the model?
- Number of Trees in a Forest:
  Is more trees in a forest better the model?
- Number of Sequential Estimators for GBM:
  How many time should we repeat sequential training?
- Grid Search vs Random Search:
  A tradeoff between efficiency and accuracy

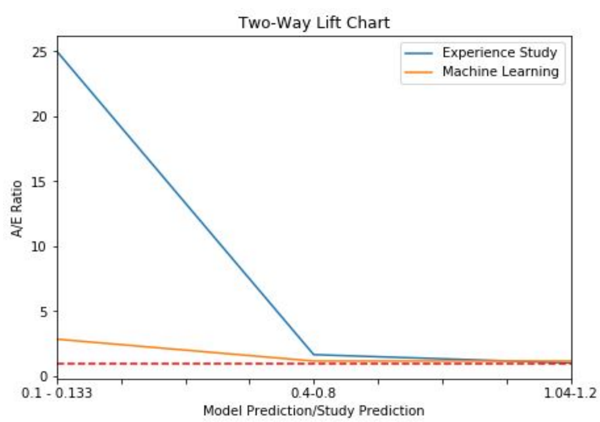
AE Ratio



It is not easy to tell which method is better here as models are compared in one-dimensional space

- Gives some sense of model performance in one dimensional space
- However, machine learning model should capture all dimensions' performance

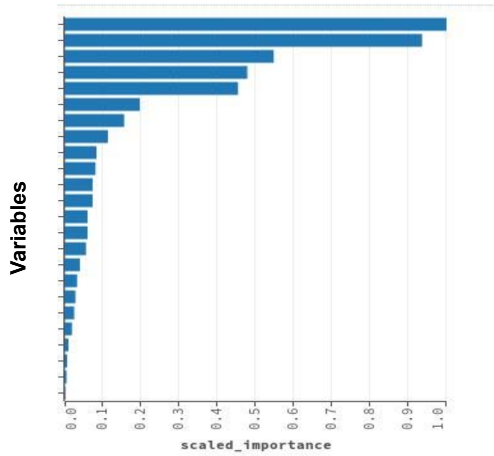
Two-Way Lift Chart



ML shows better result here as the chart consider overall dimensions

- Vertical axis is A/E ratio and horizontal axis is the ratio of machine learning model prediction to experience study prediction
- AE Ratio approach but capture more dimensions
- Better model is determined by whether the line is close to 1 throughout the range of horizontal axis

Variable Importance Chart



- Variable Importance: Calculate the relative influence of the variable in a machine learning model
- It can be used to look at the variable that has higher influence on classifying surrender policy
- We can find the variable that is not considered by traditional experience study

More on the Case Study

Method	AUC on Testing Data	Performance Increase by
Experience Study	0.70	N/A
GLM	0.81	16%
Decision Tree	0.80	12%
Random Forest	0.87	21%
GBM	0.95	29%

- Adopt GLM for model explanation while it has shown reasonable prediction power
- Random forest and GBM has shown better prediction power than decision tree
- Adopt GBM for most accurate prediction as it has the best prediction power

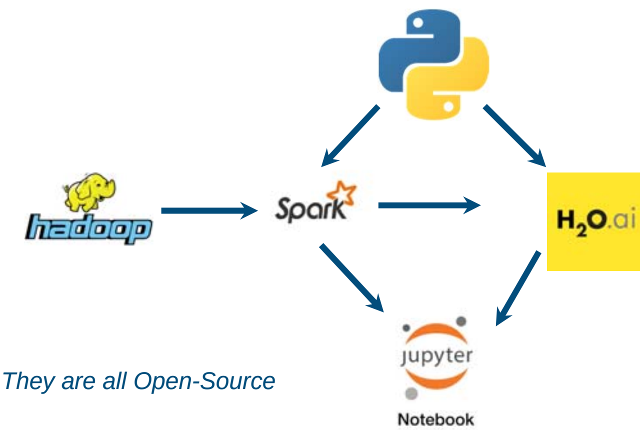
Some "Learnings" from the Study

- Machine learning suggested new dimensions not commonly looked at before in traditional experience study:

Consider the top 3 important variables → Amount of policy an agent sells affect lapse rate? → Re-examine the important variables for agents management

- Machine learning can derive a lapse function for each policy, which can be used for:
  - Lapse simulation
  - Value of customer calculation
  - Continuously monitoring lapse behavior with up-to-date data and updated model

Machine learning tool



They are all Open-Source

Machine Learning – Tool

Hadoop	Distributed data storage to store and distribute big data
Spark	Data processor for data cleaning process
H2o.ai	Package for machine learning with big data
Python	Object-oriented programming language that implement Spark and H2o.ai
Jupyter Notebook	Web-based computing interface for modelling and visualization

Q & A

SOCIETY OF ACTUARIES®

SOCIETY OF ACTUARIES®