Modeling Life Insurance Risk

Prudential Insurance Data Set

Team Name: Dataninjas

David Allen
Seung Lee

Kennesaw State University

SAS Student Symposium Forum 2016-2017

Abstract

We modeled an eight-level ordinal life insurance-risk response on a pre-cleansed and pre-normalized Prudential data set consisting of 59,381 observations and 128 predictors of which 13 were continuous, 5 discrete, and the remainder categorical. The overall objective of the project was to develop a scoring formula to simplify the life-insurance application process in order to encourage more customers to apply and, therefore, purchase life insurance. Comparison of ASEs, misclassification rates, lift, and relative parsimony led us to choose a 13-predictor logistic regression model from a pool of nine candidates. While the model, in which BMI or Body Mass Index figures prominently, is globally better than chance at classifying applicants, its misclassification error rates for response levels lower than the highest level (representing lowest insurance risk) are higher than 50 percent. The high error rates call for additional data, subject-matter expertise, and further work to refine the model.

Introduction

According to 2013-2014 U.S. census data, the life expectancy of U.S. males and females were 76.4 and 81.2 years, respectively.[1] In part because baby boomers are on the rise, life insurance companies are intensely interested in this market which was worth roughly $776 billion as of 2015.[2] Because the market is so large, companies like Prudential are eager to capture potential life insurance buyers. Sales are hampered, however, by the onerous life insurance application process which requires an enormous time investment by applicants, demands huge amounts of data, and which takes an extraordinarily long time.

Data

Prudential published a data set through Kaggle[3] containing 59,381 insurance-applicant observations and 128 predictor variables that may be used in modelling a single, eight-level insurance-risk response.[4] Due to the sensitive individual information they contain, the variables were normalized and coded with little additional information except their general nature (e.g., medical history, family history, medical keyboard, etc.) and their nominal, interval or class-variable status.

Normalized height, weight, BMI and age are also in the data set. The method used to normalize the variables is unknown. Thus, it is impossible re-transform the normalized values into their original values.

Problem

The objective of this project was to develop a simplified model for quickly and accurately binning life insurance applicants into risk classes or profiles. As a result, the typical modeling process would seek subject-matter expertise to verify the construct validity of resulting models. In truth speaking, this project was based on modeling technique and technology to obtain the objective.

---

[1] See http://www.cdc.gov/nchs/products/databriefs/db244.htm.
[2] See http://www.iii.org/fact-statistic/industry-overview.
[3] See https://www.kaggle.com/.
[4] Background information about the dataset is publicly available at https://www.kaggle.com/c/prudential-life-insurance-assessment and the limited data dictionary and download links are at https://www.kaggle.com/c/prudential-life-insurance-assessment/data.

Data Cleaning and Validation

Our primary modelling technology was SAS® Enterprise Miner™ 14.1 (EM).[5] In developing our model, we generally followed EM's SEMMA process (**S**ample, **E**xplore, **M**odify, **M**odel, and **A**ssess) reflected in the first five EM tabs shown in Fig. 1.[6] Figure 2 offers a high-altitude view of our SEMMA process.

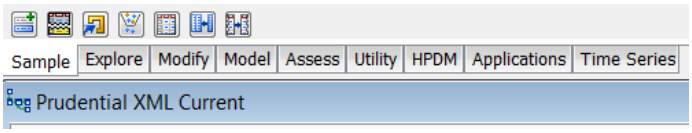


Figure 1 - EM SEMMA tabs

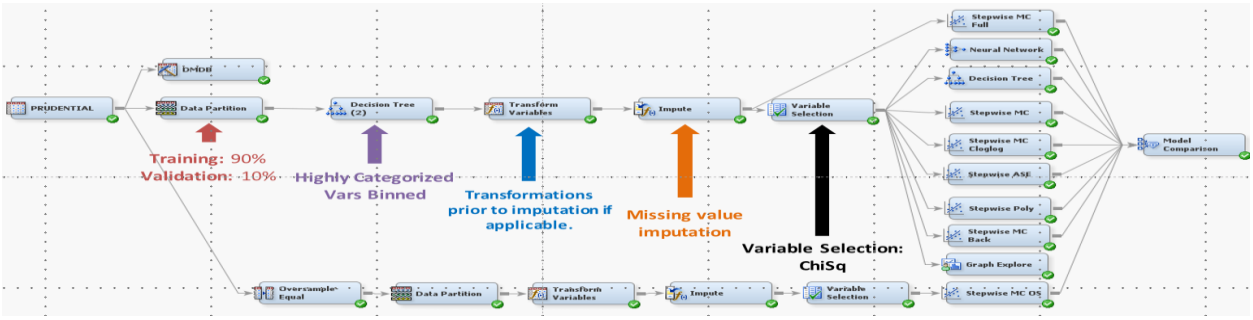


Figure 2 – EM Diagram of entire SEMMA process

Under the EM Sample tab, we first computed some initial screening statistics using the DMDB node. The DMDB output, for example, reveals that "missing" constitutes more than 50 percent of the values for interval variables Family_Hist_3 and Family_Hist_5, suggesting that they would not be useable in modeling. Several other variables (e.g., BMI, Ht, Ins_Age, and Wt) have no or relatively few missing values, making them candidate predictors. The Sample node also includes data partition, sample, and filter nodes. We partitioned the data training and validation sets. We did not use the filter node, to avoid excluding the possible impact of outliers on our models.

Under the Explore tab, we used DMDB, StatExplore, Graph Explore, and Variable Selection nodes at different stages of the process. We generated ranking candidate predictors by Chi-Square statistics, giving an early indication of which predictors might help in modeling. We also used the Explore function within the Prudential data set node to get an early picture of possible relationships among variables. During the exploration process we also inspected useable variables in EM to ensure that each variable's "level" property matched the variable type. Finally, we used a decision tree node to generate SAS discretization code for the high-dimension class variable, pdtinfo2.

Under the Modify tab, we first inserted the discretization code from the decision tree node into the SAS code property of the Transform ("Discretize pdtinfo2") node. We then impute the data set to replace missing values for interval and class variables with means and counts, respectively, but only for variables with missing value rates of less than 50%. Those with missing rates over 50% were dropped

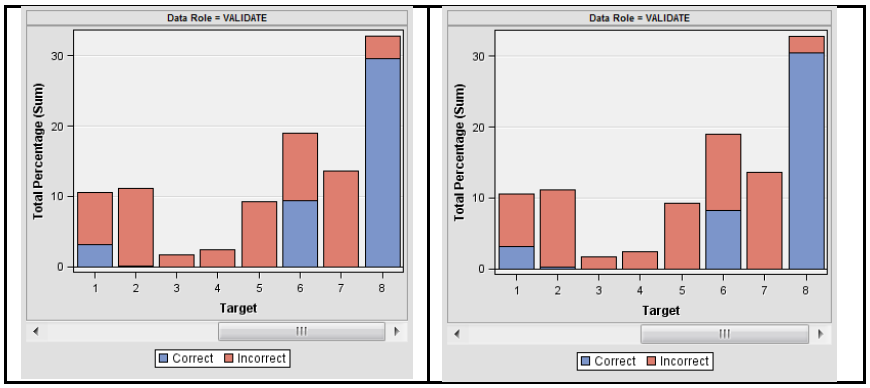
[5] We also used Base SAS 9.4 to generate some visualizations.
[6] EM also provides applications outside of the core SEMMA process under additional tabs labeled Utility, HPDM (for High-Performance Data Mining), Applications, and Time Series.

from further consideration. Following imputation, we then returned to the Explore tab, using a Variable Selection node for pre-modeling dimension reduction.

Analysis

The purpose of this analysis is to determine which model best explains the objective and what variables in the data set helps us to improve insurance applicants into risk classes or profiles. We used a Model Comparison node to compare candidate models. Based on the foregoing comparative data, we chose the "Stepwise MC" ordinal logistic regression model for its parsimony (only 13 estimates) and its competitive misclassification rate (0.5795) and ASE (0.58406). Models with better ASE, misclassification rate, and/or lift were rejected because of their complexity (e.g., decision tree) or explanatory difficulty (e.g., neural nets). Beyond aggregate misclassification error rates, with an ordinal response variable like Response, it is important to assess misclassification by response level. We decide to compare Stepwise MC and Stepwise Poly models. As a result, we found that the models are similar in their high misclassification rates with respect to all Response levels except level 8.

Table 1 — Misclassification by Response level: Stepwise MC (left panel), Stepwise Poly (right panel)



Stepwise MC was developed through a stepwise selection algorithm using misclassification error as the selection criterion.

Results and Generalizations

Normalized BMI was statistically the most important predictor (Chi-Sq = 6869, Table 9). Medical_History_4 was next, with Medical_keyword_3, Medical_History_4, and Medical_History_23, trailing the two leaders. It is important to remember that this logistic regression models log(y) on the predictors. This means that in scoring a case, the parameter estimates must be exponentiated before they are applied to the predictors. So, for example, the BMI estimate of -5.8273, when exponentiated becomes roughly .0003. Odds ratios in Table 9 (right panel) indicate the effect of changes in predictor values on the odds of an applicant reaching a higher (as we believe, lower-risk) Response level (e.g., 8 vs. 7 or 2 vs. 1). For example, with all other variables held constant, the Medical_Keyword_3 odds ratio indicates that an applicant *without* this keyword was 4.309 times more likely to move up one Response level than an applicant *with* this keyword. In contrast, BMI odds ratio of 0.0003, signifies that a 1-unit increase in BMI renders the applicant 0.003 times as likely to reach the next higher Response level than if the applicant's BMI does not change.

Table 2 – LR, Type 3 Analysis, & Odds Ratio Estimates for Stepwise MC

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
193251.154	172995.288	20255.8658	6	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
BMI	1	6868.9871	<.0001
InsuredInfo_6	1	462.8121	<.0001
Medical_History_23	1	1446.1504	<.0001
Medical_History_4	1	2809.5241	<.0001
Medical_keyword_3	1	1516.9143	<.0001
Pdtinfo4	1	1192.6376	<.0001

Odds Ratio Estimates

Effect		Point Estimate
BMI		0.003
InsuredInfo_6	1 vs 2	0.691
Medical_History_23	1 vs 3	0.486
Medical_History_4	1 vs 2	0.401
Medical_keyword_3	0 vs 1	4.309
Pdtinfo4		2.830

Working around the confidential constraints of the data to a limited extent, we can use odds ratios to conjecture desirable variable levels in terms of achieving a higher Response. As an example, a Medical_History_4 level=1 is associated with 0.405 times the likelihood of moving up one Response level compared to Medical_History_4 level=2. If this is true, then Medical_history_4 level=2 is a preferred, lower risk level than level 1.

We also hypothesized that categorizing BMI could yield informative comparisons of applicant groups. BMI values were converted into z-scores and grouped according to CDC BMI percentile indices.[7] The resulting ordinal variable was then input into the model replacing the normalized BMI variable. The subsequent adjusted model allowed us to generate predicted probabilities of response level based on BMI index given the default remaining model parameter levels. Resulting individual and cumulative plots are shown in Figures 9 below.

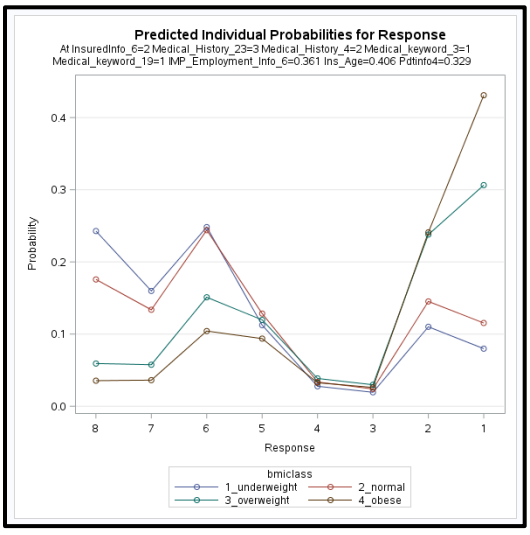


Figure 3 – Predicted Individual Response Level Probabilities by BMI class

[7] See https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/.

An obese applicant has a predicted probability of Response=1 of just over 40%, per Figure 9. By contrast, a normal BMI applicant's predicted probability Response=1 was about 12%. Interestingly, the predicted probabilities of falling into a category of response 3 or 4 was relatively identical for all BMI groups at just under 5% as seen in Figure 4. There was unavoidably some difficulty in understanding the cause for a low rate of response level 3 or 4 given the classified nature of the response and parameter levels.

We also visualized the cumulative predicted probabilities of Response levels by BMI index (Figure 10). The cumulative probability chart showed that normal BMI applicants had ≈ 70% chance of reaching the 4 highest levels (5, 6, 7, 8) in Figure 10. By contrast, overweight and obese individuals for the same response categories (5,6,7, or 8) with predicted probabilities of 40% and 28%, respectively.

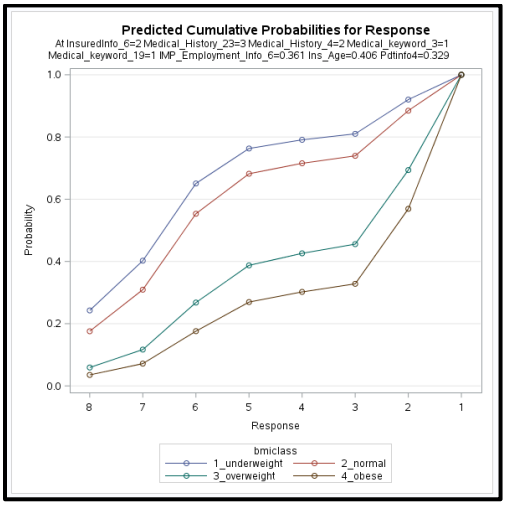


Figure 4 - Predicted Cumulative Response Level Probabilities by BMI class

Recall from that the absence of medical_keyword_3 (level=0) corresponds to an odds ratio of 4.153 of reaching a higher nominal but lower risk Response level. An applicant with normal BMI and Medical_Keyword_3=0 had a predicted probability of Response=8 of approximately 0.50 (Figure 11). By contrast, a similarly situated obese applicant's predicted probability of Response=8 is approximately 0.15. Interestingly, normal BMI crossed with Medical_keyword_3=1, produces a 0.18 predicted probability of Response=8, nearly identical to an obese individual with Medical_keyword_3=0.

In summary, despite the data set's confidentiality constraints, compelling useable information can be extracted from such observed relationships.

Suggestions for Future Studies

Further refinement of discretization routines, alteration of data partition percentages, and simply more data might serve to improve the model's accuracy. Nested binary models, one for each Response level, might be worth a look, as well as some variable interactions. The related SAS scoring code (273 lines) and EM XML diagram are available on request.

Prior to implementation of any model that would be used to make insurance-granting decisions, it would be important to get more information about the variables to ensure that their use for this

purpose is permissible under anti-discrimination laws prohibiting disparate treatment and disparate effect.

Conclusion

We modeled an eight-level ordinal life-insurance-risk response variable using ordinal logistic regression. A variety of modeling techniques generated multiple candidate models. From these candidates, we eventually chose a model built using stepwise selection and misclassification error as the selection criterion. Comparison of ASEs, misclassification rates, lift, and relative parsimony led us to choose a 13-estimate model from among the candidates. The model output indicates that while the model is much better than mere chance at predicting Response levels 1, 6, and 8 and could thus be leveraged to process client applications, the misclassification error rate with respect to all levels but 8 is a significant concern.