ASA Predictive Analytics Exam

Sample Project Report

Note to candidates: This report represents a high-quality solution. For this sample project, the creators did not consider the fact that only five hours would be available. This allows more modeling aspects to be covered than are possible in five hours. However, in recognition of the time pressure, some of the tables and graphs are not presented in as reader-friendly a form as would be possible with more time. However, extra time was taken to set the parameters of rpart.plot to make the tree easier to read when copied to Word. This is not necessary, but if the tree is hard to read in the Word version, it is important that the Rmd file produces a readable version.

To: My Supervisor

From: PA Candidate

Date: Today

Title: Drivers of Term Life Insurance Purchasing

Executive Summary

Critical questions about our term insurance product are (1) who is likely to purchase it? and (2) for those who do, who is likely to purchase larger policies? These questions are important because the answers can assist us in making cost-effective marketing decisions. While internal data might help answer these questions, it is more important to learn about individuals who are not yet our customers. To learn about such individuals, our department obtained information from the Survey of Consumer Finances. Applying analytical techniques, we were able to determine relationships that can help guide the marketing department in its efforts to determine likely purchasers.

With regard to who is likely to purchase term insurance, among those with low income (less than $25,500), those not yet retirement age (less than 62) and with low charitable giving (below $1300) are unlikely to purchase term insurance but others with low income are. Also, among those with high income (greater than $25,500), those outside prime working years (older than 62 but younger than 28) generally do not purchase term insurance, nor do those inside prime working years with charitable giving in excess of $169,000.

For the size of the policy purchased, we found that the key factors leading to larger amounts of insurance are:

- Being male;
- Being younger;

- Having a spouse or partner;
- Having more education;
- Having a larger household;
- Having a higher income;
- Giving more to charity; and
- Having a smaller age difference between the two spouses/partners.

Some of the data used in this analysis may be inappropriate to use due to legal, privacy, or customer concerns. These concerns should be addressed before the model is put to use.

Data Exploration, Preparation, and Cleaning

I was provided data from the Survey of Consumer Finances, a nationally representative sample that contains extensive information on assets, liabilities, income, and demographic characteristics of those sampled. The sample used for this analysis contains data on 500 households with positive incomes that were interviewed in the 2004 survey. It is important to note that the information is about the person who completed the survey. That need not be the individual who is the owner of the insurance policy or the life that is insured.

This data was collected by the U.S. Government, which has fewer restrictions on what may be collected. Before implementing any marketing policies based on this analysis, our legal department should be consulted to learn if we are legally barred from obtaining some of this information. There is also a risk that customers will react poorly to being asked for some of this information as they will know that is being tied back to them. Should you decide that some of the variables used in the final model are not legal or not feasible, let me know and the models can be rerun without them.

There are two target variables to consider. One is FACE, which is the amount of the death benefit. A value of zero means that no insurance was owned. The other is a derived variable, TERM_FLAG, which was set equal to 1 if FACE is positive and 0 otherwise.

Nine predictor variables extracted from the survey were deemed to have a possible relationship with at least one of the target variables. The list of predictor variables, as well as descriptions of what they represent, is provided in Appendix A.

The following table is a statistical summary of the two target and nine predictor variables from the original dataset.

```
      FACE             TERM_FLAG         GENDER            AGE            MARSTAT        EDUCATION
 Min.   :       0   Min.   :0.00   Min.   :0.000   Min.   :20.00   Min.   :0.00   Min.   : 2.00
 1st Qu.:       0   1st Qu.:0.00   1st Qu.:1.000   1st Qu.:37.00   1st Qu.:0.00   1st Qu.:12.00
 Median :   10000   Median :1.00   Median :1.000   Median :47.00   Median :1.00   Median :14.00
 Mean   :  411170   Mean   :0.55   Mean   :0.826   Mean   :47.16   Mean   :0.79   Mean   :14.06
 3rd Qu.:  200000   3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:58.00   3rd Qu.:1.00   3rd Qu.:16.00
 Max.   :14000000   Max.   :1.00   Max.   :1.000   Max.   :85.00   Max.   :2.00   Max.   :17.00
      SAGE         SEDUCATION         NUMHH           INCOME            CHARITY
 Min.   : 0.0   Min.   : 0.00   Min.   :1.00   Min.   :     260   Min.   :      0
 1st Qu.: 0.0   1st Qu.: 0.00   1st Qu.:2.00   1st Qu.:   28000   1st Qu.:      0
 Median :39.0   Median :12.00   Median :2.00   Median :   54000   Median :    500
 Mean   :33.4   Mean   :10.02   Mean   :2.87   Mean   :  321022   Mean   :  34089
 3rd Qu.:51.0   3rd Qu.:16.00   3rd Qu.:4.00   3rd Qu.:  106000   3rd Qu.:   3000
 Max.   :78.0   Max.   :17.00   Max.   :9.00   Max.   :75000000   Max.   :9010000
```

The following observations stand out:

- 55% of the individuals bought term insurance. The sample is nicely balanced between those who did and did not purchase term insurance.
- The age range of 20 to 85 is reasonable. This is the current age, not the age at which the policy was purchased. Spouse age (SAGE) of zero likely indicates no spouse.
- Education ranges from 2 to 17. The small values may be questionable. Spouse education (SEDUCATION) of zero likely indicates no spouse.
- The number in the household has a reasonable range of 1 to 9.
- The minimum income of 260 seems small and the maximum of 75,000,000 seems large, so there may be outliers. The same applies to the maximum of charitable giving.
- There are no missing values per se (though, as noted, a value of zero may mean the field doesn't apply).

Histograms of the three economic variables (plotting only positive values) shows extreme right skewness, which can often be remedied by taking logarithms. The following are histograms of their logarithms, which support use of this transformation. This will be done, with zero values for FACE and CHARITY kept at zero.

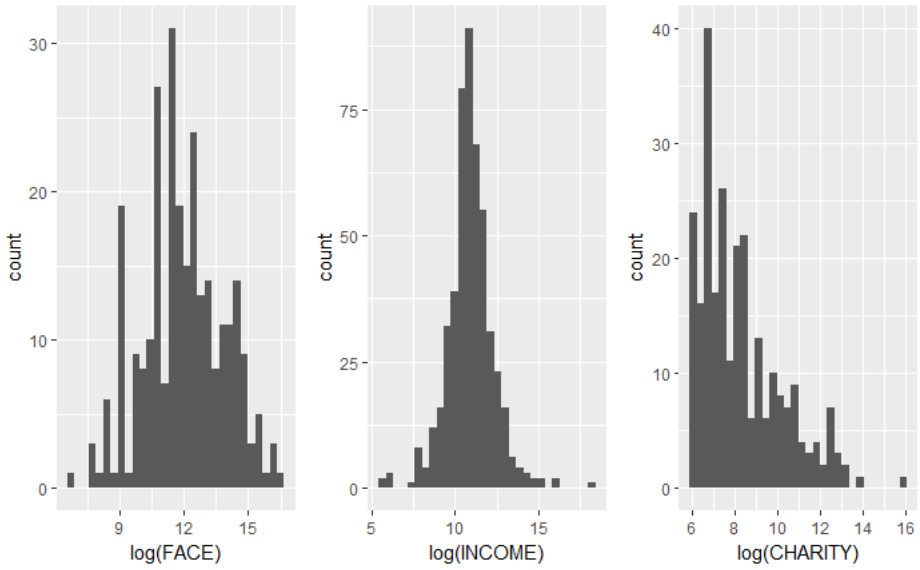


Figure 1: Histograms of the logarithms of the three economic variables.

The data were checked to ensure that all who have MARSTAT = 0 have spouse values of zero, which turns out to be the case. There were a few cases with low values of education. The 20 records with values of EDUCATION and SEDUCATION (where there is a spouse) less than or equal to six were removed.

The one value of CHARITY above 1,000,000 was deemed outliers and removed. Large values should normally first be investigated to see if there is a coding error or other explanation. I was not able to do that and thought it best to remove them so they will not affect the analysis.

The final observation was to note that there were only 28 values where MARSTAT = 2. It was decided to change those values to 1 so there are only cases of those with or without a spouse/partner.

The final step was to save the data in two files. One is the full (remaining) set of observations to be used to model the TERM_FLAG variable. The second is the subset where FACE is positive, to be used to model the log(FACE) variable.

Feature Selection

Based on discussions with marketing, two features were added. When buying term life insurance, FACE is believed to be strongly related to INCOME, such that it may be more appropriate to use the ratio FACE/INCOME as the target variable. This new variable was created. Second, a possible driver for buying term insurance is a large difference in ages between the respondent and spouse. A large difference may indicate a greater need for insurance should the older spouse die (noting that we do not know if it is in fact the older spouse who is the insured). To test this hypothesis, a new variable, the absolute difference between AGE and SAGE (where SAGE is positive) was created.

I also explored two additional ways to generate additional features. First, I used Principal Components Analysis to combine the six variables that each had limited predictive power into a single feature. When added to the four original variables that had significant predictive power, there was no gain in predictive power, so this feature was not added. Second, I applied clustering to see if the two education variables could be combined. *K*-means clustering indicated there might be value in mapping them to three qualitative levels, those with no spouse/partner, those where the combined education years is less than 29 and those where it is greater than or equal to 29. However, again, this new feature did not provide any increase in predictive power, and so will not be used.

Model Selection and Validation

TERM_FLAG Target Variable

The variable that indicates purchase of insurance is TERM_FLAG, which is 0 if not purchased and 1 if term insurance is purchased. There are ten available predictor variables. They are the nine original variables plus AGEdiff, the absolute difference in the spouse ages. Our goal is to have a model that is easily interpretable. Two possible model choices are classification trees and logistic regression. I chose to focus on classification trees for three reasons:

- The model is more flexible. Regression requires a linear relationship between the features and the target, trees do not.
- The model is easier to interpret because it characterizes potential customers by looking at where they stand in regard to selected variables rather than using an equation.
- It performs feature selection as part of the model building process.

These combined attributes should yield a model that can be easily used by marketing to develop its strategies.

The first step was to set up training and validation data sets. A random sampling approach was used with each observation having an 80% chance of being in the training set. 384 of the 479 (80.2%) observations ended up in the training set.

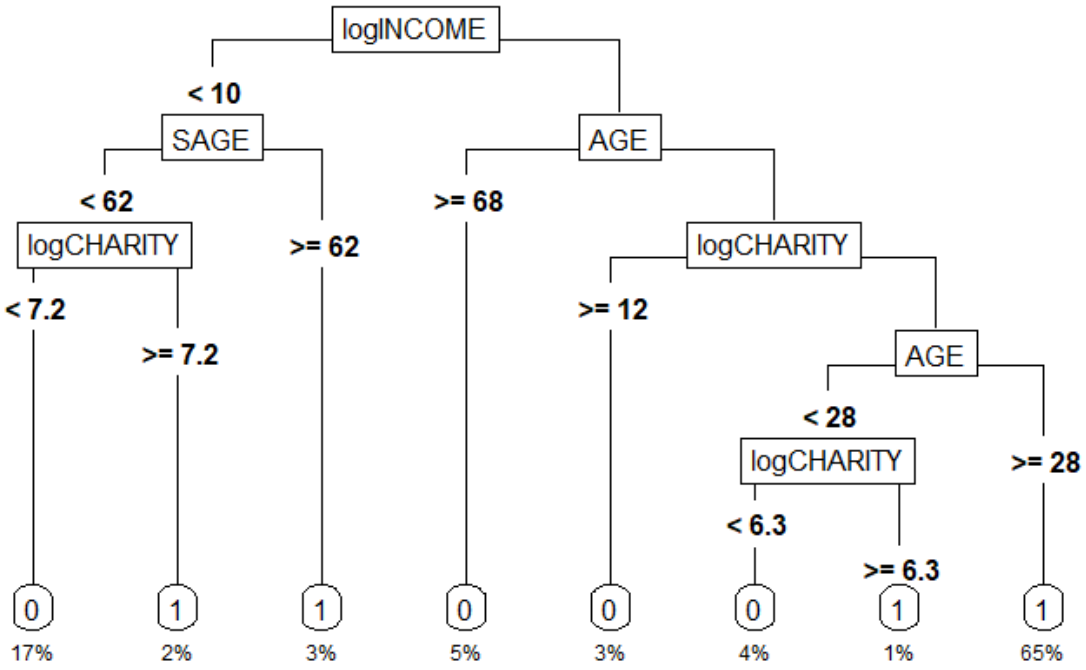
The first tree examined used recursive binary splitting, the gini index, and cost complexity pruning. When the resulting tree was evaluated against the validation set, it correctly classified 61 of the 95 observations (64%), with the results shown below. A similar result was obtained using cross validation.

		Actual	
		0	1
Predicted	0	18	12
	1	22	43

I next tried the random forest approach. Values of mtry (the number of features to include in each tree in the random forest) from 1 to 7 were tried. There was a slight improvement, with 65 of 95 (68%) classified correctly. However, random forest results are not as easy to interpret or communicate and so the slight improvement in predictive ability is offset by a loss in ease of model use. However, a variable importance analysis (see the table below) showed that AGE, logCHARITY, logINCOME, and SAGE were most important. These are the same four features used in the previous trees, confirming that the simpler model is appropriately identifying the most important features.

Variable	Importance
AGE	100
logCHARITY	92
logINCOME	77
SAGE	58
SEDUCATTION	48
NUMHH	32
GENDER	21
AGEdiff	15
MARSTAT	7
EDUCATION	0

Given that it is easier to explain a single tree, I decided to apply the first method to the full dataset. The following tree was obtained: A node of “1” indicates that the prediction is the household has purchased term insurance, and the percentages are the proportion of observations that end at that leaf.



FACE Target Variable

There are only 270 observations where the FACE variable is positive. They were randomly split with 217 (80.4%) in the training set. With a continuous target variable, either regression trees or a generalized linear model (GLM) could be used. I have elected to focus on the GLM approach. A problem with trees is that every potential insured who has the same node is predicted to have the same FACE value. With a small dataset, we saw in the previous section that only a few nodes are likely to occur. Selecting a GLM is more difficult here. We have three potential target variables (FACE, logFACE, and FACEratio), a variety of link and distribution functions, and a variety of feature selection approaches. In each case, combinations are used that ensure predictions are consistent. For example, FACE must be positive, so a gamma distribution makes sense. Using a log link ensures the predicted mean is always positive. The root mean squared error (RMSE) will be used to compare the various models.

One complication when fitting a regression model, that was not present when using trees, is the relationship between marital status (MARSTAT) and any feature, such as spousal age, that is zero when MARSTAT = 0. If the spousal feature is retained but MARSTAT dropped, the difference between 0 and 1 will be the same as between 1 and 2, yet 0 has a special meaning. Hence, if any of those variables are retained, MARSTAT should also be retained.

Model 1: Predict logFACE using an identity link and the normal distribution

Note that this model is actually an ordinary least squares model. However, we will use a GLM approach to be consistent with subsequent models. Backward selection using AIC indicates features should be dropped in the following order: SEDUCATION, SAGE, and MARSTAT. However, from the above discussion, because AGEdiff is retained we should keep MARSTAT. Eliminating only the two variables produces an RMSE of 787,109.

Model 2: Predict FACE using a log link and the gamma distribution

There were convergence issues. The best model obtained had an RMSE of 1,029,868.

Model 3: Predict FACEratio using a log link and the gamma distribution

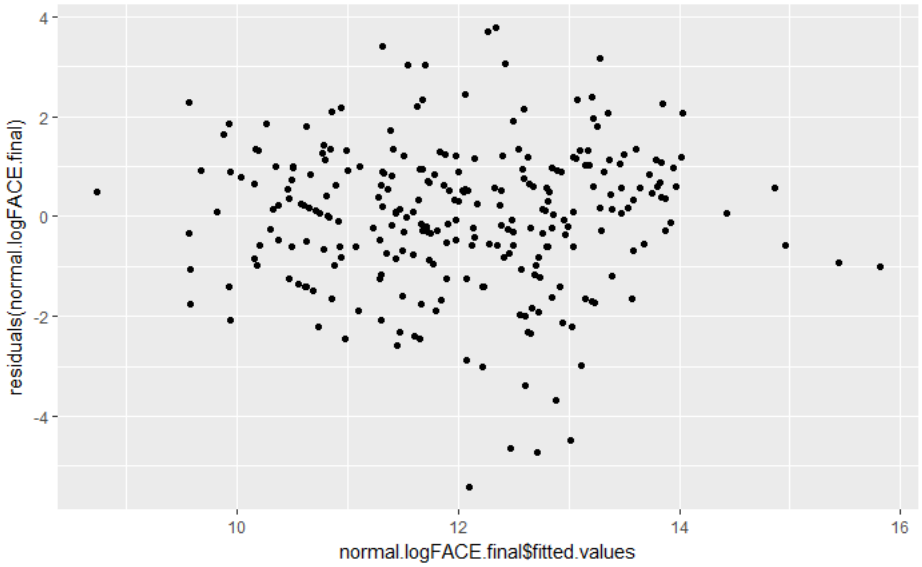
There were convergence issues. The best model obtained had an RMSE of 868,380.

It appears that Model 1 is the best one to use. It may be possible to do better by employing a method that more directly addresses the bias-variance tradeoff. The method selected here is cross-validation using the lasso. The lasso is preferred because it allows some features to be eliminated, which may lead to a simpler model for marketing to use. The result was the elimination of AGE and SAGE. The RMSE was 1,029,869.

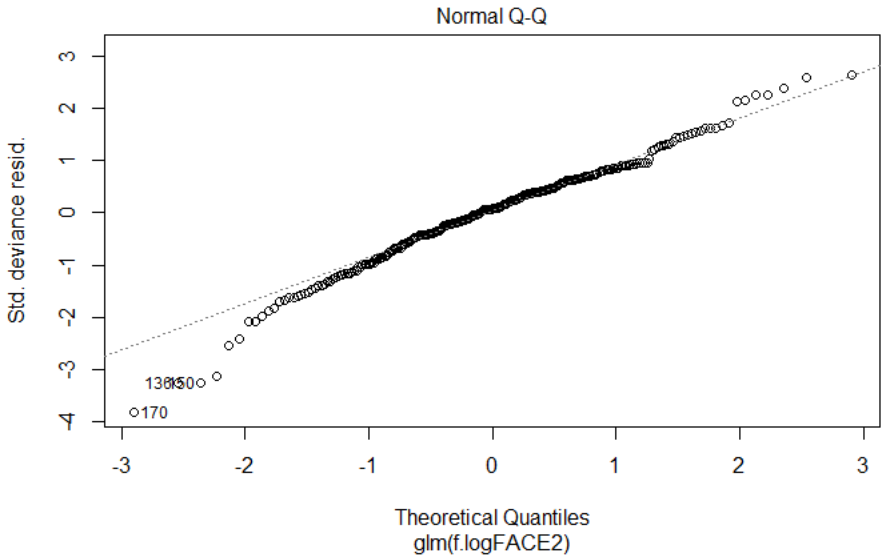
In the end, the simple ordinary least squares model with two variables removed was selected. It was fit to the full dataset, with the following results.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.476886   1.049588   4.265 2.79e-05 ***
GENDER       0.763147   0.360535   2.117 0.035231 *
AGE         -0.012702   0.008206  -1.548 0.122863
MARSTAT      0.068088   0.342067   0.199 0.842380
EDUCATION    0.169014   0.049166   3.438 0.000683 ***
NUMHH        0.229631   0.072202   3.180 0.001648 **
logINCOME    0.339385   0.081221   4.179 4.01e-05 ***
logCHARITY   0.118339   0.027491   4.305 2.37e-05 ***
AGEdiff     -0.047119   0.026305  -1.791 0.074411 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The following diagnostic plots were obtained to ensure that the assumptions were (approximately) met.



Residual plot of GLM.



Q-Q plot of GLM.

The residual plot looks good, with a consistent vertical spread from left to right. The left end of the q-q plot appears to fit less well, indicating the model may not do as well at predicting smaller face amounts.

Findings and Recommendations

Separate investigations were done to develop models for predicting who will buy term insurance and for predicting how much coverage will be purchased by those who do. Because the data used were by household units, the results are actually predictive of someone in a given household purchasing insurance; the model does not identify who within the household owns the policy or whose life is insured.

For who might buy term insurance, it is not surprising that the main indicator is income, as seen in the decision tree. More specifically:

- Among those with low income (less than $25,500), those not yet retirement age (less than 62) and with low charitable giving (below $1300) are unlikely to purchase term insurance but others with low income are.
- Among those with high income (greater than $25,500), those outside prime working years (older than 62 but younger than 28) generally do not purchase term insurance, nor do those inside prime working years with charitable giving in excess of $169,000.

For predicting the amount purchased, an equation was developed. It starts with a baseline of $88. This seems small, but each factor presented below is a multiplicative adjustment. The following table indicates those adjustments.

Factor	Adjustment
Gender	Multiply by 2.15 if male
Age	Multiply by 0.987 for each year of age
Marital Status	Multiply by 1.07 if there is a spouse/partner
Education	Multiply by 1.184 for each year of education
Number in household	Multiply by 1.26 for each household member
Income	Multiply by income raised to the 0.339 power
Charitable giving	Multiply by giving to the 0.118 power.
Difference in ages	Multiply by 0.954 for each year of difference

We can program this formula for your use. It is presented here to give you an idea of how the factors influence the amount purchased. Males purchase over twice as much, the amount goes down (slightly) with age, there is a slight bump for having a spouse partner, education (18% more purchased for each year of education) has a strong effect as does the number in the household (26% increase per member), and there is a modest reduction as the difference in ages increases. The income and charitable giving effects are harder to interpret, but it suffices to say there is a strong relationship to each.

As noted earlier, some of these variables may not be practical, suitable, or allowable. Let me know if such issues arise.

Appendix

Appendix A – Data Dictionary

The following table describes the eleven variables that were used in this analysis.

Variable Name	Description
FACE	Amount of the death benefit
TERM_FLAG	0 if no term insurance purchased, 1 if term insurance purchased
GENDER	0 = Female, 1 = Male
AGE	Age in years of the respondent
MARSTAT	Marital status of the respondent: 1 = married, 2 = not married, but living with a partner, 0 = all other
EDUCATION	Number of years of education of the respondent
SAGE	Age of spouse/partner
SEDUCATION	Number of years of education of the spouse/partner
NUMHH	Number of household members
INCOME	Annual household income
CHARITY	Charitable contributions

Note: A value of zero could indicate that the response is 0 or could indicate a missing or inapplicable value.