# Kaggle Modeling Competition for Analytics Edge Course

*Ed Kashmarek*

## Situation

In 2014 I took an online course in data analytics called the Analytics Edge offered by MIT through edX. In week 7 we had a modeling competition on Kaggle. The idea was to predict whether a person was happy or not based on answers to questions using logistic regression. The contestants were allowed to submit as many models as they wanted. The models were trained on a given training set, tested on a given test set and scored for the competition on another test set that was not available to the contestants. There were over 1,500 entries in the contest.

## Task

We were asked to build a model using any model type we learned in the course, including generalized linear models (GLM), Classification and Regression Trees models (CART) and random forest models (RF).

## Action

I built several models using all different types. One of my best models was a GLM model (see below). The model had 13 independent binary variables. The variables were household status (domestic partners, married, single) and the answers to ten questions. The questions chosen were those that had at least one answer that was significant at the 0.05 level in a previous model I ran, which had questions included based on my own intuition about happiness.

## Result

The results are as follows:
Model  AUC (Area Under ROC (Receiver Operator Characteristic) Curve) on training set: 0.7419257
Model accuracy (true positives + true negatives as % of all observations) on training set: 0.6909784
Model $R^2$ (strength of model compared to baseline model with no variables) on training set: 0.1745963
Model RMSE (root mean squared error) on training set: 0.4505914
Model AIC (measure of quality to compare models) on training set: 4722
Model AUC (Area Under ROC Curve) on test set: 0.71529
Independent variables:

- Significant at 0.001 level:
  - Are you an optimist or a pessimist? (Optimist)
  - Are you an optimist or a pessimist? (Pessimist)
  - Do you feel like you are normal? (No)
  - Are you in over your head? (No)
  - Are you in over your head? (Yes)
  - Are you more successful than high school friends? (No)
- Significant at 0.01 level:
  - Household status married (no kids)
  - Household status married (with kids)
  - Does life have a purpose? (Yes)
  - Do you live alone? (Yes)
  - Does your life feel adventurous? (No)
  - Did you accomplish anything inspiring in 2013? (No)
- Significant at 0.05 level:
  - Do you feel like you are normal? (Yes)
  - Do you enjoy giving or receiving more? (Giving)

Interpretation of Results

The model I used was a generalized linear model for logistical regression. The goal was to predict whether someone was happy (1) or not happy (0) using certain predictor variables. The model AUC was 74.2%, meaning that given a random person who was actually happy and a random person who was actually not happy, the model would predict the correct classification of each person 74.2% of the time on the training set. The accuracy of the model was 69.1%, meaning that the model predicted the true positives (those actually happy) and the true negatives (those actually not happy) 69.1% of the time when tested against all observations in the training set, as opposed to random observations as with AUC. The model $R^2$ on the training set was 0.17, meaning the model is a 17% improvement over a baseline model with no independent variables. Put another way, the model explains 17% of the variation in the dependent variable (Happy=1). Although the $R^2$ is quite low, good models can have low $R^2$, especially in logistical regression. The model RMSE on the training set was 0.45, meaning the average difference between the predicted value and the actual value was 0.45. Given that the values for happiness range from 0 to 1, this is a fairly large RMSE. However, considering the independent variables are binary, this large RMSE makes sense, since happiness can change significantly depending on the answer to a particular question. All of the models I ran had an RMSE in this range. The model AIC on the training set was 4722, which has no real interpretation other than to compare it against the AIC of other models, with the goal being to have the lowest AIC among a set of models. The model AUC on the test set was 71.5%, slightly less than the AUC of 74.2% on the training set. Thus, when the model was tested on data other than the data used to build the model, the model did not do quite as well. Still, an AUC of 71.5% is pretty good.

Analysis

As for predicting happiness, the analysis is as follows. In this model, by far the most significant variable is answering “No” to the question “Are you in over your head right now?” The significance level for this variable is 0.001, which is the probability that we incorrectly reject the null hypothesis ($H_o$: the coefficient is not significantly different from zero) when in fact the null hypothesis is true. The p-value, or the probability of obtaining at least as extreme results as the estimated coefficient, given that the null hypothesis is true, for this variable is basically zero. Since the p-value is less than the significance level, we can reject the null hypothesis with a great deal of confidence that we will not make an incorrect rejection. How confident are we that we are correctly rejecting the null hypothesis? If the significance level, or probability of an incorrect rejection, is 0.001, then the probability of a correct rejection is 99.9%. Since the p-value is less than 0.001, we can be more than 99.9% confident that we will not be making an incorrect rejection. This means that we are very confident that the true coefficient is in fact not zero, and the best estimate of the coefficient is 0.6368321.

So what does that number mean? The probability of the dependent variable Happy=1 is as follows:

Prob(Happy=1) = $1/(1+e^{-L})$

Where: L is the logit and is equal to $B_0+B_1X_1+B_2X_2+\ldots+B_KX_K$

$B_0$ is the intercept

$B_1$ is the coefficient estimate for $X_1$, the first independent variable

$B_K$ is the coefficient estimate for $X_K$, the $K^{th}$ independent variable

If a person answers “No” to this question, the value of the logit rises by 0.6368321, all else being equal. When the logit rises, the value of $e^{-L}$, or $1/ e^L$, declines, which means the denominator declines and, hence, the value $1/(1+e^{-L})$ rises. Thus, the larger the coefficient, the bigger the impact on the probability that the person is happy, while the smaller the coefficient, the smaller the impact.

The bigger the difference between the coefficients of the binary answers for each question, the bigger will be the difference in the probability of the person being happy based on which answer is given for the question. Since the coefficient for the answer “No” to this question is large and positive, while the coefficient for the answer “Yes” to this question is fairly large and negative, there is a very big difference in the probability of being happy based on the answer to the question “Are you in over your head?” If you are in over your head you are much more likely to be unhappy, while you are much more likely to be happy if you are not in over your head, all else being equal.

The next most significant variable is saying “No” to the question “Are you more successful than your high school friends?” The p-value for this variable is 0.000319 and the coefficient for this variable is -0.4473253. If a person says “No” to this question, the probability of happiness decreases. Interestingly, even if a person says “Yes” to this question, the probability of happiness decreases slightly. However, the p-value is very high, so a “Yes” answer to this question is insignificant, meaning it does not materially affect happiness one way or the other.

Here are the most significant variables with answers (in order of significance) and their impact on the probability of happiness:

| | |
|---|---|
| Are you in over your head right now? No | (increases probability of Happy=1) |
| Are you more successful than high school friends? No | (decreases probability of Happy=1) |
| Do you feel like you are normal? No | (decreases probability of Happy=1) |
| Are you an optimist or pessimist? Pessimist | (decreases probability of Happy=1) |
| Are you an optimist or pessimist? Optimist | (increases probability of Happy=1) |
| Are you in over your head right now? Yes | (decreases probability of Happy=1) |
| Household status: married with no kids | (increases probability of Happy=1) |
| Did you accomplish anything inspiring in 2013? No | (decreases probability of Happy=1) |
| Does life have a purpose? Yes | (increases probability of Happy=1) |
| Does your life feel adventurous? No | (decreases probability of Happy=1) |
| Household status: married with kids | (increases probability of Happy=1) |
| Do you live alone? Yes | (decreases probability of Happy=1) |
| Do you feel like you are normal? Yes | (increases probability of Happy=1) |
| Do you enjoy giving or receiving more? Giving | (increases probability of Happy=1) |

Here are the most significant variables with answers (in order of effect on happiness based on their coefficients) and their impact on happiness:

| | |
|---|---|
| Are you in over your head right now? No | (increases probability of Happy=1) |
| Household status: married with no kids | (increases probability of Happy=1) |
| Household status: married with kids | (increases probability of Happy=1) |
| Are you an optimist or pessimist? Optimist | (increases probability of Happy=1) |
| Does life have a purpose? Yes | (increases probability of Happy=1) |
| Do you enjoy giving or receiving more? Giving | (increases probability of Happy=1) |
| Do you feel like you are normal? Yes | (increases probability of Happy=1) |
| Do you feel like you are normal? No | (decreases probability of Happy=1) |
| Does your life feel adventurous? No | (decreases probability of Happy=1) |
| Are you in over your head right now? Yes | (decreases probability of Happy=1) |
| Did you accomplish anything inspiring in 2013? No | (decreases probability of Happy=1) |
| Are you an optimist or pessimist? Pessimist | (decreases probability of Happy=1) |
| Are you more successful than high school friends? No | (decreases probability of Happy=1) |
| Do you live alone? Yes | (decreases probability of Happy=1) |

Thus, the biggest increase to happiness is not being in over your head, while the biggest decrease to happiness is being single. The exact effect on the probability of happiness for these variables cannot be determined since it will be different for each person based on their answers to the other questions.

As alluded to above, we can analyze this another way. If we look at the absolute difference between the coefficients of the binary answers for each question, we can get a better read on how happiness differs based on answers to particular questions. For example, if a person answers “No” to the question “Are you in over your head?”, the coefficient is about 0.64. Yet if the answer is “Yes”, the coefficient is about -0.38. Thus, the absolute difference in the coefficients based on the answer to this question is 1.02. Another example, if a person says “Yes” to the question “Does life have a purpose?”, the coefficient is about 0.36, but a “No” answer yields a coefficient of about 0.18, so the absolute difference is 0.54. But there is a catch here. Since the “No” answer is not significant at any level, we can treat that coefficient as being zero for absolute comparison purposes. So, for this question, the absolute difference would be 0.36.

Similar analysis of the different questions yields the following absolute differences (AD) in coefficients:

| | |
|---|---|
| Are you in over your head right now? (Yes/No) | 1.02 AD in coefficients |
| Are you an optimist or pessimist? | 0.85 AD in coefficients |
| Do you feel like you are normal? (Yes/No) | 0.62 AD in coefficients |
| Do you live alone? (Yes/No) | 0.46 AD in coefficients |
| Are you more successful than high school friends? (Yes/No) | 0.44 AD in coefficients |
| Did you accomplish anything inspiring in 2013? (Yes/No) | 0.43 AD in coefficients |
| Does your life feel adventurous? (Yes/No) | 0.38 AD in coefficients |
| Does life have a purpose? (Yes/No) | 0.36 AD in coefficients |
| Do you enjoy giving or receiving more? (Giving/Receiving) | 0.27 AD in coefficients |
| Married couples (kids/no kids) | 0.08 AD in coefficients |
| Single people (kids/no kids) | no AD in coefficients |
| Hardship in life result of…(circumstances/decisions) | no AD in coefficients |
| Domestic partners (kids/no kids) | no AD in coefficients |

Since the AD in coefficients is largest for the question “Are you in over your head right now?”, it stands to reason that the biggest difference in the probability of happiness for any given person would be how they answered this question, all else being equal. Similarly, the answers to the bottom three questions would have no difference in the probability of happiness. But again, we cannot determine the exact impact since the impact of the answer to each question will depend on the answers to other questions.

Thus, according to this method, the best way to increase your happiness is to not be in over your head, while the smallest difference is having no kids if you are married. There is no difference for the bottom three variables since there is no absolute difference in their coefficients.

## Conclusion

Either way you analyze this data, it is clear that not being in over your head adds the most to happiness. Thus, according to this analysis, if you want to be happy, start by reducing your workload or stress level, whether at work, at home, in relationships or in any aspect of your life. Besides making you more happy, it will likely improve your health as well. In addition, finding a significant other (not in the statistical sense!) can make you smile more!

Here, in layman’s terms, are the best ways to increase your happiness, according to this model:

Don’t get in over your head
Be more optimistic about life
Feel normal
Live with someone
Be more successful than your high school friends

Accomplish something inspiring
Add some adventure to your life
Find a purpose for your life
Give more
Don’t have kids if you get married

## Contest

As for the contest, this model did not place very well, in the bottom half if I recall. However, competition was fierce, and there were many experienced data scientists and computer programmers in the field. I did build and submit another model using the random forest approach using all variables except Year of Birth, Gender, Income and Political Party, with a node size of 200 and 5000 trees as the parameters. The result was an AUC of 0.77004 and 252nd place, in the top 15%! Not bad for my first contest! Unfortunately, that model was much more difficult to interpret and the number of variables used would have meant many more hours spent on this write-up. In any case, it was a wonderful learning experience!

## Model

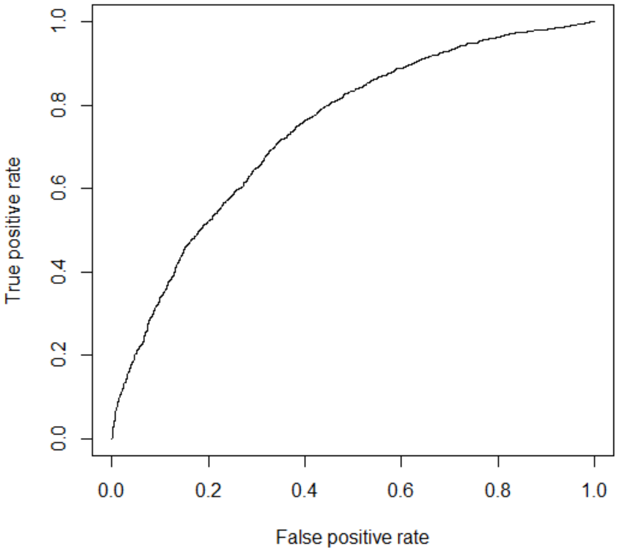
```
Dependent variable: Happy (0=not happy, 1=happy)=
Independent variables:
Coefficients:
                                                        Estimate     Std. Error    z value  Pr(>|z|)
(Intercept)                                             0.1109327    0.1600169     0.693    0.488149
HouseholdStatusDomestic Partners (no kids)              0.2913717    0.2485771     1.172    0.241134
HouseholdStatusDomestic Partners (w/kids)               0.0712622    0.4010942     0.178    0.858983
HouseholdStatusMarried (no kids)                        0.5780066    0.1762983     3.279    0.001043 **
HouseholdStatusMarried (w/kids)                         0.5036937    0.1583916     3.180    0.001472 **
HouseholdStatusSingle (no kids)                         0.1012181    0.1513950     0.669    0.503770
HouseholdStatusSingle (w/kids)                          -0.1232999   0.2280441     -0.541   0.588725
Q98869No Does life have a purpose?                      -0.1847786   0.1343118     -1.376   0.168901
Q98869Yes Does life have a purpose?                     0.3579497    0.1120343     3.195    0.001398 **
Q99716No Do you live alone?                             -0.1561405   0.1207403     -1.293   0.195944
Q99716Yes Do you live alone?                            -0.4698187   0.1669748     -2.814   0.004897 **
Q101162Optimist Optimist or a pessimist?                 0.4181664   0.1204185     3.473    0.000515 ***
Q101162Pessimist Optimist or a pessimist?               -0.4448784   0.1269869     -3.503   0.000459 ***
Q102289No Does your life feel adventurous?              -0.3752732   0.1176097     -3.191   0.001419 **
Q102289Yes Does your life feel adventurous?             0.1778531    0.1303931     1.364    0.172575
Q107869No Do you feel like you're "normal"?             -0.3671912   0.1042925     -3.521   0.000430 ***
Q107869Yes Do you feel like you're "normal"?            0.2532014    0.1058980     2.391    0.016803 *
Q115899Circumstances Hardship in your life result of…   -0.1359873   0.1034837     -1.314   0.188815
Q115899Decisions Hardship in your life result of…        0.0628855   0.1017141     0.618    0.536405
Q118237No Are you "in over-your-head" right now?        0.6368321    0.1123767     5.667    1.45e-08 ***
Q118237Yes Are you "in over-your-head" right now?       -0.3779400   0.1094403     -3.453   0.000554 ***
Q119334No Accomplish anything inspiring in 2013?        -0.4323308   0.1340253     -3.226   0.001256 **
Q119334Yes Accomplish anything inspiring in 2013?       -0.0002378   0.1328456     -0.002   0.998572
Q119650Giving Enjoy more: giving or receiving?          0.2719985    0.1329506     2.046    0.040770 *
Q119650Receiving Enjoy more: giving or receiving?       0.1871295    0.1501451     1.246    0.212645
Q120014No More successful than high-school friends?     -0.4473253   0.1242857     -3.599   0.000319 ***
Q120014Yes More successful than high-school friends?    -0.0228024   0.1158843     -0.197   0.844008
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 5391.6  on 3934  degrees of freedom
Residual deviance: 4667.9  on 3908  degrees of freedom
AIC: 4721.9
```

## R Code for Model

```
train=read.csv("train.csv")
nrow(train)
train=na.omit(train)
nrow(train)
train$UserID=NULL
train$votes=NULL
train$YOB=NULL
test=read.csv("test.csv")
nrow(test)
glmmod9=glm(Happy~HouseholdStatus+Q98869+Q99716+Q101162+Q102289+Q107869+Q115899
+Q118237+Q119334+Q119650+Q120014,data=train,family=binomial)
summary(glmmod9)
predTrain9=predict(glmmod9,type="response")
str(predTrain9)
table(train$Happy,predTrain9>0.5)
library(ROCR)
ROCRPred9=prediction(predTrain9,train$Happy)
auc=as.numeric(performance(ROCRPred9,"auc")@y.values)
auc
SSE=sum((train$Happy-predTrain9)^2)
SST=sum((train$Happy-mean(train$Happy))^2)
R2=1-SSE/SST
R2
RMSE=sqrt(SSE/nrow(train))
RMSE
predTest9=predict(glmmod9,newdata=test,type="response")
table(predTest9>0.5)
table(train$Happy)
ROCRPerf9=performance(ROCRPred9,"tpr","fpr")
plot(ROCRPerf9)
```

## ROC Curve


True positive rate
False positive rate