Modeling Mortgage Defaults - Part 1: Logistic Regression

Objective

To explore the model creation process using real mortgage data to predict a 6-year probability of default.

Logistic regressions are useful in predicting outcomes for many economic, social-science, and health applications give prior a prior status/set of characteristics; the probability of passing a standardized test given a child's profile; the likelihood of contracting a disease given one's risk exposures; the likelihood of retaining a customer given their prior behavior.

We present an abridged version of the actual model development process followed to predict 6-year mortgage default probability. We've attempted to simplify as far as possible while remaining illustrative. We create a model using on a few variables but discuss how a larger investigation would incorporate and assess other variables. Often times we present only glimpses into larger analytic discussions.

We touch on the following aspects:

- data sources - implications, biases;
- variable selection and transformation decision-making;
- model choice;
- model estimation – interpretation, assessment of fit, and sensitivity analysis;
- incorporating real-world feedback;

Questions are of vital importance, especially those relating to the core objective. Often through the data exploration / model creation process we obtain insights better reformulate our initial question, and arrive at our 'true' objective in a more direction fashion.

Why are we interested in a 6 year default rate? Would another tenor suffice? For simplicity we pursue this unquestioningly. Similarly, what is default? How do we define it? (We'll answer this below).

A helpful and general analytic framework is described by Roger Peng in 'The Art of Data Science'. He describes talks about cycling problems and subproblems through the following core activities:

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models
4. Interpreting the results
5. Communicating the results

Data Source

The data source is part of Fannie Mae's (FNMA) Single Family Loan portfolio (http://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html). The data contains only 30-year, fixed-rate mortgages, originated under specified programs and guidelines. This data was release publicly by FNMA in 2011. It contains origination and performance data going back to 1999. It contains 10 million loans, with over 15 years of monthly performance history.

We can divide the loans into testing and training data. There is usually a tradeoff between variability in your estimators (training size too small), and variability in your performance statistic (testing size too small). With such a large data set this is less of a concern. We start with an 80/20 split of training to testing data, sampled randomly. We don't need to use all the training data for each model estimation. By estimating a similar model once using all the training data we can determine how large a sample we would need to maintain similar precision in loadings and errors. We would then continue to use data samples of that size in the exploratory phase. For the final model we would want to use the entire training set.

For simplicity and practical purposes we randomly sample 10,000 loans from FNMA's 2005-Q1 acquisitions. Extrapolating heterogeneity from a largely homogenous population is problematic, and we will see some of the shortfalls of only using 2005-Q1 acquisitions in subsequent modules, but it is illustrative nonetheless.

FNMA does not originate the mortgages for its portfolio, but instead buys them from thousands of nation-wide originators that underwrite to FNMA's guidelines. FNMA's guidelines are extremely comprehensive and concern many of the hundreds and thousands of data points necessary to review during the mortgage underwriting process. However, idiosyncracies can still exist between originators.

Some data-related questions that we need to consider while building our predictive model include:

- How applicable is this dataset to other mortgage portfolios we hope to analyze?
- FNMA's acquisition guidelines have changed over time, how re-usable is a model estimated on one set of underwriting guidelines for predicting defaults on loans originated under another?
- Guidelines aren't the only things that have changed over time, credit scoring methodologies (e.g. how FICOs are calculated) have changed over time. Similarly, asset appraisals (the V in LTV) change over time or are inflated. Won't current housing price bubbles affect the equity a borrower has in his property and ultimately their propensity to default? How can we control for this?

We also notice that in any given quarter FNMA acquires both newly originated and seasoned mortgages. Our 2005-Q1 acquisition sample includes some mortgages originated as early as 1999. The acquisition process necessarily induces selection bias (that is infact what we are try to estimate). However, if a loan is already 6 years old and we include it in our sample aren't we biasing our 6-year default rate down since our portfolio doesn't include, and FNMA wouldn't have purchased, that same loan had it defaulted?

Economic data is often messy and not exactly what you want. Yes, FNMA's acquisition criteria has changed over time, but by how much? We reviewed the changes in underwriting and they often cover a level of nuance not captured in our data tape. Calculation methods for variables like FICO and LTV have changed over time. We unfortunately have no insight into how FICO has changed. The V in LTV is a subjective concept. We can try to capture the variability in its calculation through other means, for example by including an home-price index covariate (often a lagging-indicator), to help identify pricing corrections. To address the positive performance selection bias from only being able to choose performing seasoned loans, we can perhaps filter our training sample to have a no greater than a given time-to-acquisition. Determining this level is qualitative, and may include factors as practical as the number of loans remaining after filtering by time-to-acquisition.

Model Choice

Logistic models are often used for regression problems where the predictors are continuous and/or categorical variables. They are powerful models, easily estimatable, with accessible and intuitive interpretations of coefficients and outputs. The logistic regression is based around the idea of a bernoulli distribution, but extended across the range of values each variable can take. Consistency is enforced across proximate continuous covariates due to the functional form fitted.

Where β and X are vectors of the loadings and the covariates respectively, the functional form of the logistic regression is

$$p(Y|\beta) = \frac{e^{\beta X}}{1 + e^{\beta X}}$$

. The function has an 'S' shape (if you want to think of it in one dimension), taking on values between 0 and 1.

The logistic regression has a useful interpretation. Each coefficient represents the increase in log-odds of the response for a unit increase in the covariate. If we exponentiate the coefficient, it tells us the increases in the odds, i.e. $\frac{prob(Y|\beta)}{1-prob(Y|\beta)}$, of the response, from increasing just that covariate. Notice that the change in odds with respect to a covariate (without interaction terms) is constant. This is similar to a linear model where the change with respect to a covariate is constant.

The logistic regression is essentially a one-period model. We measure an observation's characteristics (covariates) at a starting time, return at an ending time, and observe the response (i.e. dependent variable) for each observation

- It does not matter when the response happened, just that it happened
- The model does not permit censored or incomplete observations, ie. the outcome for each observation must be {response occurred, response did not occur}.
- Borrower characteristics cannot change over time
- The outcome of a logistic regression is a single number, the probability of response.

Caveats / Concerns

All models have requirements/assumptions, violation of some are more tolerable than others. The logistic regression requires:

- Independent observations
- Larger datasets – Maximum likelihood methods usually require more observations per variable to be estimated correctly

To the extent we are missing variables, or the model is not close to the 'right' one we will have bias in our model. For simplicity we purposely select only a few variables, but describe a more general process to variable selection.

Variable Selection and Transformation

Variable selection is constrained by data availability, but also by creativity. We trade between significant economically grounded variables and blind significance testing (ex. forwards/backwards variable selection methods) for supplemental variables.

Variable selection is consistently a trade-off between over-fitting and parsimonious generalization. A helpful process is to abstract from actual data and outline the high level factors determining the response - we could divide variables of interest into the following components - consumer credit history, current credit profile, property type/characteristics, macroeconomic indicators.

A strong understanding of context can often help with the creativity side of variable selection. In our data, we see that California and Florida had large asset price bubbles and some of the highest default rates in the country. We can create an indicator variable identifying whether, after controlling for other variables, knowing a loan is in CA or FL tells you anything more about its default probability.

Covariate Creativity

To capture the impact of macroeconomic indicators in the logistic regression context we must be slightly more creative. Microeconomic theory tells us broadly that agents will strategically default when the loan balance is higher than property value, i.e. the borrower has negative equity. Moreover, lower house prices also indicate a weakening housing market, economy, and credit performance generally. This suggests default probability is related to home price movement, specifically downward home price changes. We investigate a covariate containing the lowest relative property value over the estimation period.

This covariate comes close to violating a principle of a forecasting model to not use future information to predict the future. In creating this variable we must be careful to not take the lowest relative property price up to loan default (this would be using observation-specific future default information). Instead, by using non-specific home-price information of the lowest relative price over the estimation interval we are still able to capture the intent of variable without unfairly reaching into the (past's) future. This becomes clearer when we think about using the model. We can (and should) have views about future states of the economy (e.g. what will happen in 2,4,6 years), and this should affect our default estimates on a loan, but how could we estimate what the lowest relative HPI will be prior to a loan's default? What model would we use to determine when the loan will default and would therefore be able to say we've found the minimum of the relative HPI movement? We don't have a means of knowing the failure time of a loan and so we cannot at origination supply this covariate. However, we can supply the lowest relative HPI value over the prediction interval (e.g. 6 years), by means of another model, our own views, etc.

Model Estimation - interpretation, assessment of fit, and sensitivity analysis

We load the data and estimate the model

```r
Logistic1 <- glm(RESPONSE_LOG ~ OCLTV + DTI + CSCORE_B + CA.FL +
    SATO + HPI_CHANGE, data = LogTrain, family = "binomial")

summary(Logistic1)
```

```
##
## Call:
## glm(formula = RESPONSE_LOG ~ OCLTV + DTI + CSCORE_B + CA.FL +
##     SATO + HPI_CHANGE, family = "binomial", data = LogTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7358  -0.2289  -0.1668  -0.1139   3.5618
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.277984   1.268336   0.219   0.8265
## OCLTV        0.047806   0.006978   6.851 7.35e-12 ***
## DTI          0.008663   0.006894   1.257   0.2089
## CSCORE_B    -0.009079   0.001417  -6.405 1.50e-10 ***
## CA.FL       -0.366680   0.247394  -1.482   0.1383
## SATO         0.544422   0.261218   2.084   0.0371 *
## HPI_CHANGE  -2.389507   0.674933  -3.540   0.0004 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1591.9  on 7999  degrees of freedom
## Residual deviance: 1459.9  on 7993  degrees of freedom
## AIC: 1473.9
##
## Number of Fisher Scoring iterations: 7
```

This shows us that DTI is not significant at the 95% confidence interval, neither is knowing whether the loan was originated in CA-FL. We re-estimate the model without these covariates and re-verify that all our coefficients are statistically significant.

```
Logistic2 <- glm(RESPONSE_LOG ~ OCLTV + CSCORE_B + SATO + HPI_CHANGE,
    data = LogTrain, family = "binomial")

coef(Logistic2)
```

```
##   (Intercept)        OCLTV     CSCORE_B         SATO   HPI_CHANGE
## -0.004893471  0.049097377 -0.009059916  0.539355632 -1.833598619
```

```
exp(coef(Logistic2) * c(0, 10, 25, 0.25, -0.1))
```

```
## (Intercept)       OCLTV    CSCORE_B        SATO  HPI_CHANGE
##    1.000000    1.633906    0.797321    1.144352    1.201247
```

Interpretation of Coefficients

The coefficients represent an increase in the log-odds of default. If we multiply the coefficients by reasonable unit changes and exponentiate we can find the increase in default odds for a 10-point increase in OCLTV, a 25-point increase in FICO, a 25bp increase in SATO, or a 10% drop in lowest-relative-HPI.

Assessment of Fit

'All models are wrong, but some are useful' - George Box

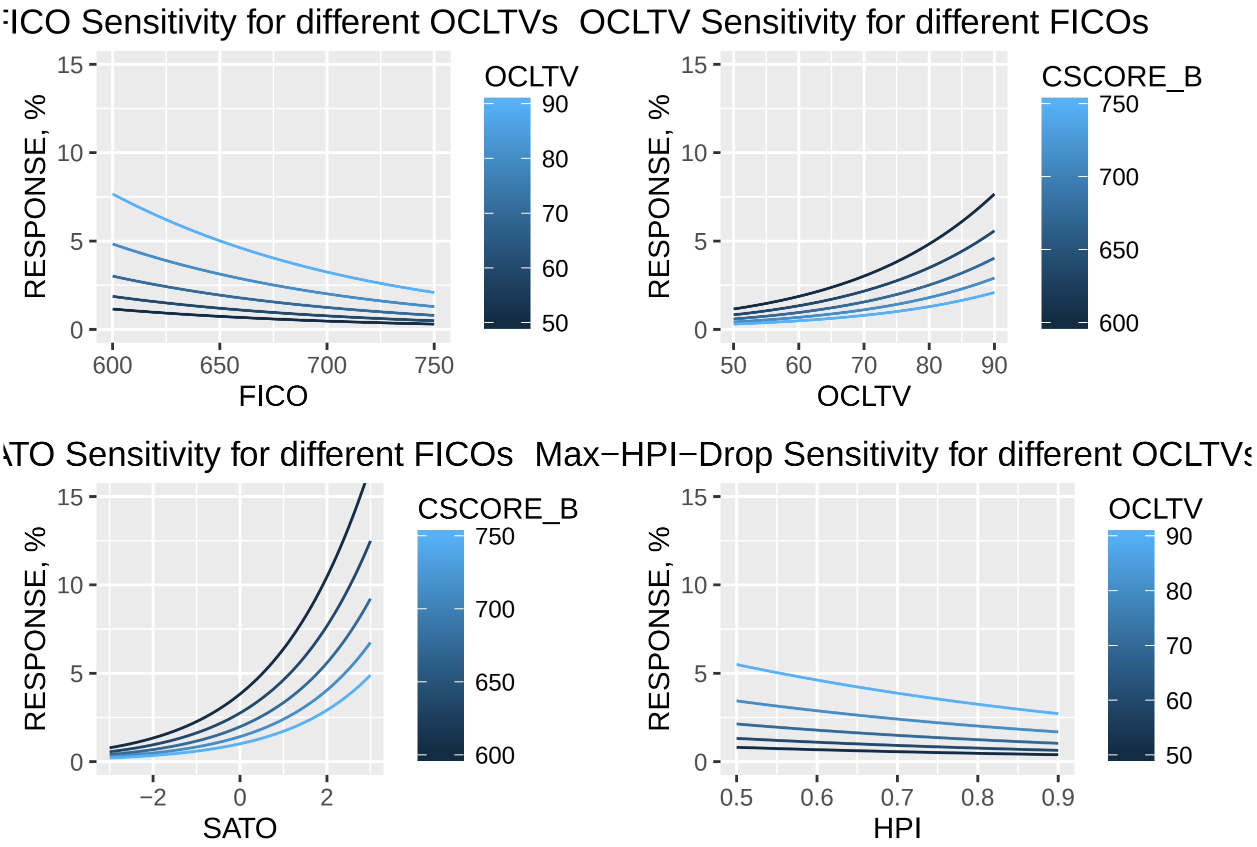
The primary measure of a model is performance on out-of-sample data. We present below some general measures of fit. We look at model vs. empirical, weighted and unweighted, default rates since this ultimately is the desired output from the model. We reperform the model vs. empirical comparison for different strata of our testing sample - high & low FICO, high lowest-relative-HPI, low lowest-relative-HPI, etc, etc. Holes in the model, or areas of large deviation from empirical data should be clearly communicated.

```
## The raw unweighted default rate is  1.75%
##  The modeled unweighted default rate is  2.0001%
##  The raw weighted-average default rate is 1.3898%
##  The modeled weighted-average default rate is  1.964%
```

Other measures of fit for the logistic regression include the adjusted R^2 statistic and the Kolmogorov-Smirnov statistic.

Sensitivity Analysis

We show model sensitivities to FICO, OCLTV, SATO and Lowest Relative HPI. Where loan characteristics are not specified we use baseline values of FICO=700, OCLTV=80, SATO=0.05.



Incorporating Real-World Feedback

SATO has significant statistical explanatory power. We desire covariates that have statistical significance as well economic grounding, and avoid those that contradict economic sense. After speaking with numerous originators about thier internal processes we learn that SATO is often not used by orignators to discriminate credit quality. Originators will often charge higher SATOs, and consumers are willing to pay them, for softer benefits such as swift and painless mortgage closings.

After an industry survey we determine that approximately half of current originators have use SATO exclusively as a non-credit measure. We consider dampening the effect of SATO on our model. Dampening SATO is to introduce some bias into the loadings of each covariate The univariate sensitivities above will be biased as re-estimated they ‘explain’ more of response that SATO used to explain.

However, if we make the exogenous decision that a free (i.e. unconstrained) SATO impact is not appropriate or accurate we can dampen its effect. We would likely need to feel confident that this non-credit SATO regime operated during . The contribution of each covariate may be biased, but the overall probability could be closer to the truth if our belief of dampened SATO loadings is correct. Below we estimate a final model with a dampened SATO loading of 50%.

```
freeSATOcoeff <- Logistic2$coefficients["SATO"]

Logistic3 <- glm(RESPONSE_LOG ~ OCLTV + CSCORE_B + offset(0.5 *
    freeSATOcoeff * SATO) + HPI_CHANGE, data = LogTrain, family = "binomial")

summary(Logistic3)
```

```
##
```

```
## Call:
## glm(formula = RESPONSE_LOG ~ OCLTV + CSCORE_B + offset(0.5 *
##     freeSATOcoeff * SATO) + HPI_CHANGE, family = "binomial",
##     data = LogTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6815  -0.2306  -0.1696  -0.1164   3.5645
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.223802   1.127435   0.199 0.842650
## OCLTV        0.049697   0.006891   7.212 5.51e-13 ***
## CSCORE_B    -0.009359   0.001383  -6.767 1.32e-11 ***
## HPI_CHANGE  -1.876108   0.544654  -3.445 0.000572 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1584.2  on 7999  degrees of freedom
## Residual deviance: 1464.6  on 7996  degrees of freedom
## AIC: 1472.6
##
## Number of Fisher Scoring iterations: 7
```