Survival Analysis of SBA data:

Motivation:

Survival analysis gives more detailed information about how the default risk of a loan varies over time. With binary classification, we estimated the probability that a given loan *ever* defaults. With a hazard model, we are able to estimate the probability that a loan defaults between any two points of time in its life.

Model Choice:

There exist many specialized Cox models that assume a particular form of the baseline hazard function. The Cox Proportional Hazards Model does not have this requirement. We can see this in the following description of the partial maximum likelihood procedure used to estimate the paramaters of the Cox PH model:

The form of the cox model is:

$$h(t) = h_0(t)exp(\beta^T X)$$

Suppose there are r observed death times in the data (all distinct), and that t_j is a death time in the set of possible death times: $R = \{t_1, t_2, ..., t_r\}$.

Then the conditional probability that an individual dies at time t_j given t_j is a time of death in the set R:

$$\frac{P(\text{individual with feature vector } X^{(j)} \text{ dies at } t_j)}{P(\text{one death at } t_j)}$$

$$= \frac{P(T = t_j | X^{(j)}, T \geq t_j)}{P(T = t_j | X^{(k_0)}, T \geq t_j) \cup P(T = t_j | X^{(k_1)}, T \geq t_j) \cup ... P(T = t_j | X^{(k_q)}, T \geq t_j)}$$

Where $k_0, ..., k_q$ correspond to the indices of observations with event times greater than or equal to t_j. Since the probabilities in the denominator are *assumed to be conditionally independent*, the denominator can be expressed as a sum of probabilities. Converting the above to continuous time, we get:

$$= \frac{\lim_{\delta \to 0} \frac{P(T < t_j + \delta | X^{(j)}, T \geq t_j)}{\delta}}{\sum_{i=k_0}^{k_q} \lim_{\delta \to 0} \frac{P(T < t_j + \delta | X^{(i)}, T \geq t_j)}{\delta}}$$

$$= \frac{h_j(t_j)}{\sum_{i=k_0}^{k_q} h_i(t_j)} = \frac{h_0(t_j)exp(\beta^T X^{(j)})}{\sum_{i=k_0}^{k_q} h_0(t_j)exp(\beta^T X^{(i)})} = \frac{exp(\beta^T X^{(j)})}{\sum_{i=k_0}^{k_q} exp(\beta^T X^{(i)})}$$

And we can see that the contribution of any observation to the likelihood function will not be dependent on h_0. □

Data

Roughly 95% of loans in the training data set had a term of 20 years. We decided that considering loans with the same term was more appropriate for this analysis (84,949 loans).

Within the training data, about 86% of loans were right censored (term did not expire in window, and did not default), about 7% of loans were paid off (term expired in window), and about 7% of loans defaulted within the window (figure 1).

Polynomial terms up to *degree five* were added for all numeric variables. Our intention was to conduct feature selection during model fitting (through regularization).

All numeric variables were centered to 0, and scaled by standard deviation.

Missing values were set to 0 and an missing value indicator feature was added for each original variable.

Including expanded categorical variables, polynomials, and missing value dummies, the data had 201 features.

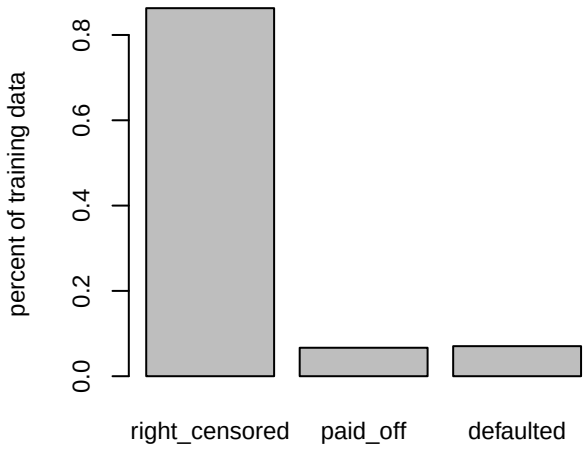


Figure 1: Loans in training data by status

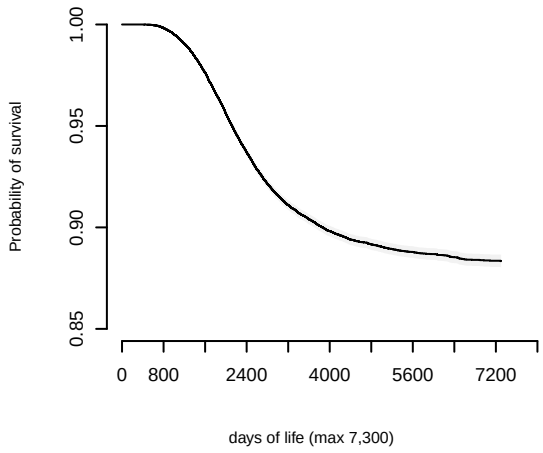
Kaplan-Meier Survival Curve

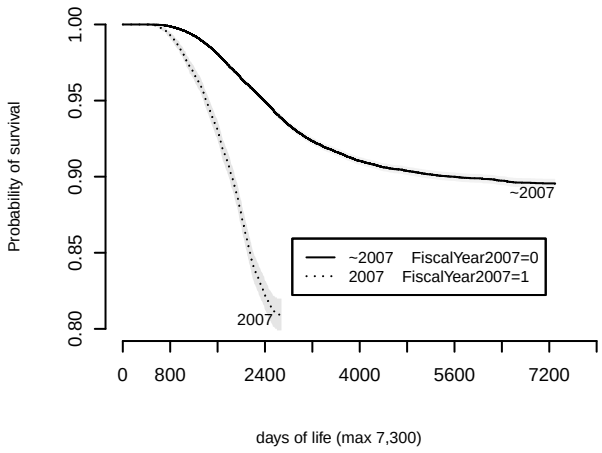
A Kaplan-Meier curve is a non-parametric estimate of the survival function, $S(t) = P(T > t)$, defined as:

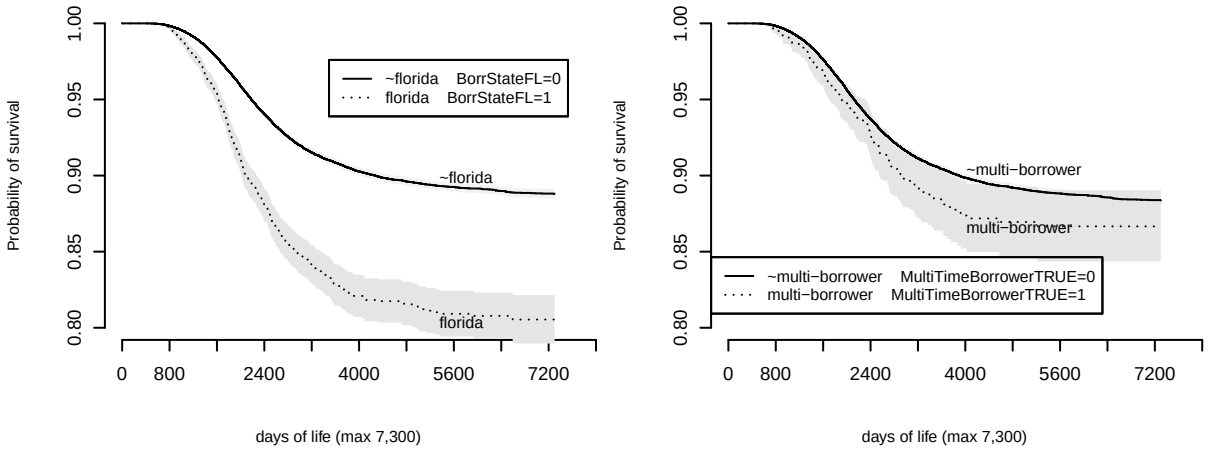
$$\hat{S(t)} = \prod_{t_i \leq t} [1 - \frac{d_i}{n_i}]$$

Where $\{t_1, ..., t_r\}$ are the death times of observations in the data, $\{d_1, ..., d_r\}$ are the number of deaths that occur at those times, and $\{n_1, ..., n_r\}$ are the number of observations remaining in the at-risk population just before those times.

For expository purposes I have included several examples of the estimated survival function (conditioned on variables such as a particular year, state, or status, as well as the general survival curve for our population).







Penalized Cox Proportional Hazards Model

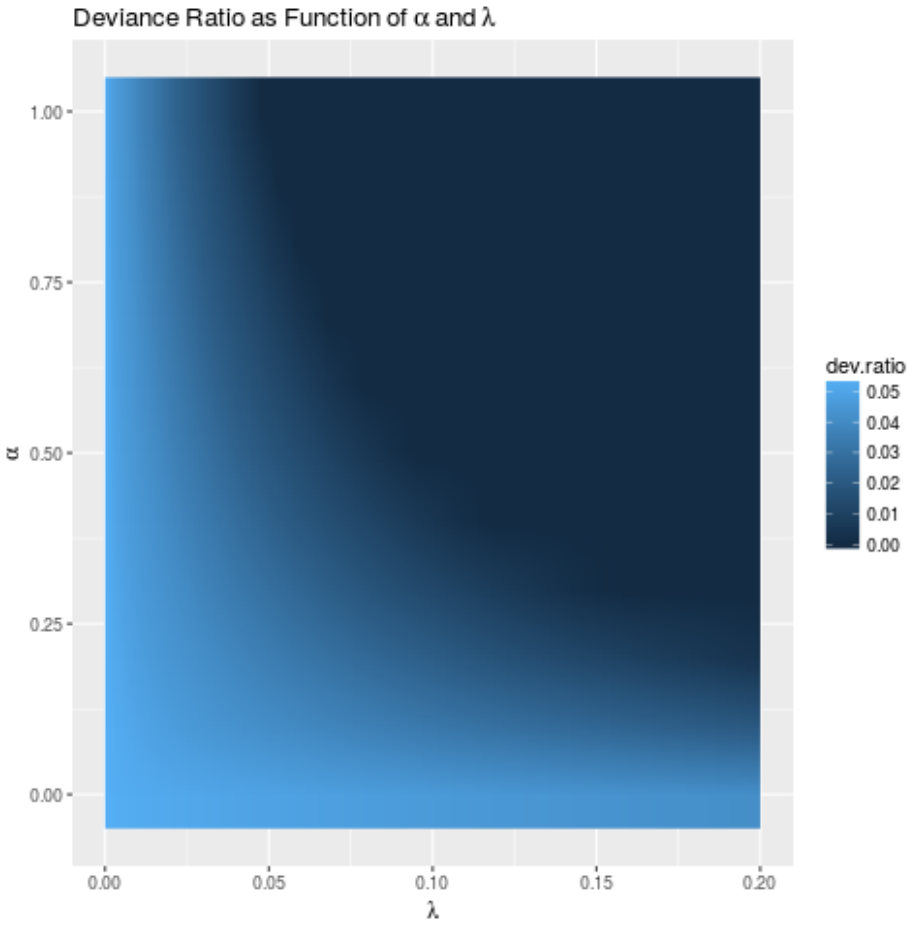
For the purpose of feature selection, we fit a series of penalized Cox models to the training data.

We used an elastic net penalty– a penalty term that is a linear combination of the l_1 and l_2 penalties.

$$\lambda[(1-\alpha)||\beta||_2 + \alpha||\beta||_1]$$

We fit models varying α and λ to maximize a goodness of fit measure used by the `glmnet` package in R, defined as follows:

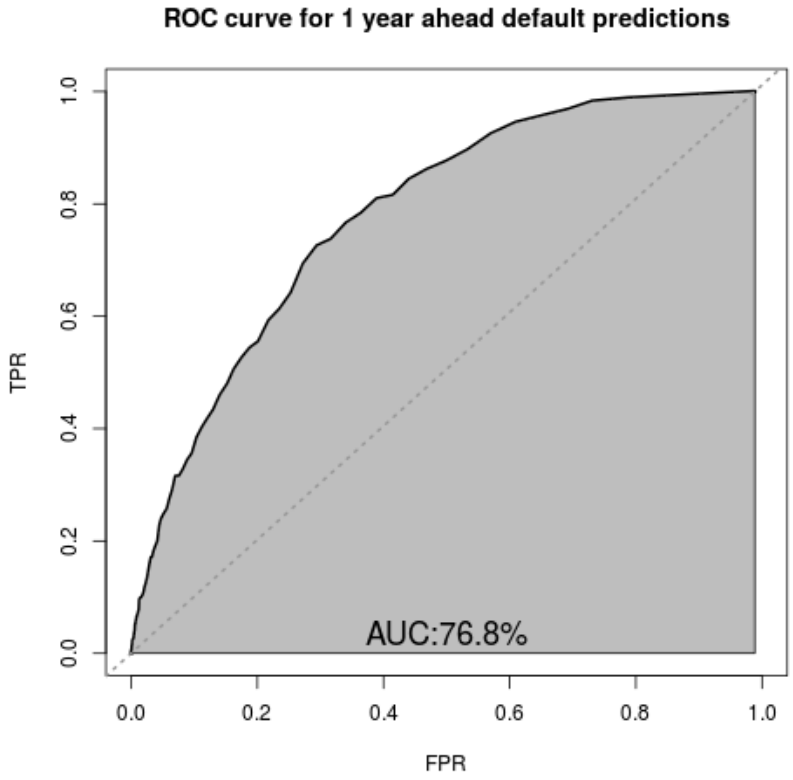
> The fraction of (null) deviance explained. The deviance calculations incorporate weights if present in the model. The deviance is defined to be 2(loglike_sat - loglike), where loglike_sat is the log-likelihood for the saturated model (a model with a free parameter per observation). Null deviance is defined to be 2(loglike_sat -loglike(Null)); The NULL model refers to the 0 model. dev.ratio=1-deviance/nulldev.

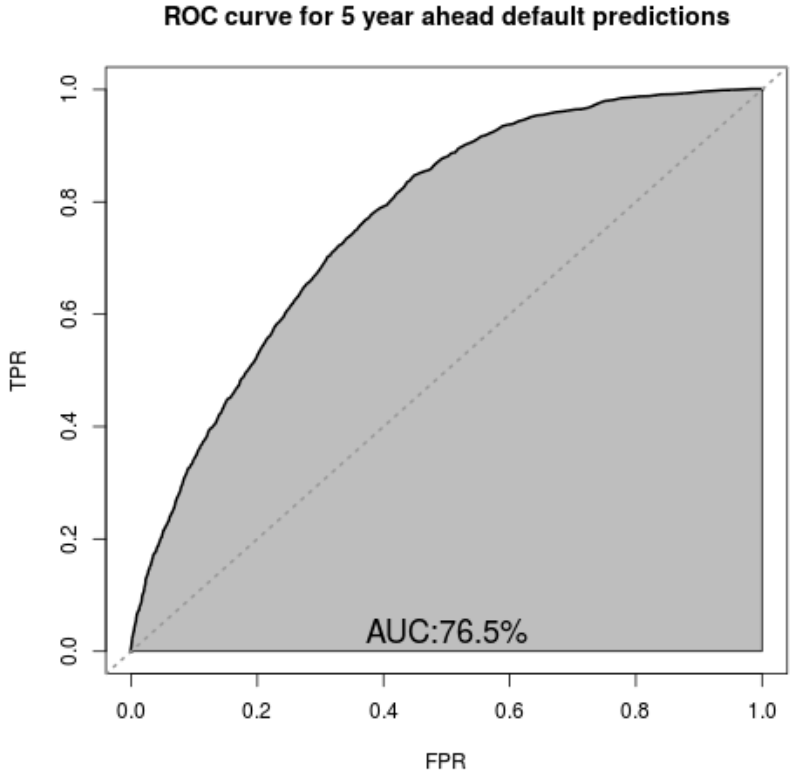


The best model, in terms of deviance ratio had a value of λ very close to 0, and α very close to 0 (the ridge penalty). Ninety-seven variables of the original 201 had non-zero coefficients.

```
## [1] "Best model by max dev ratio:"
## [1] "alpha: 0"
## [1] "lambda: 9.99999999998225e-06"
## [1] "dev.ratio:0.0538879085090161"
## [1] "number of vars selected: 97"
```

One Year and Five Year Prediction of Defaults


ROC curve for 1 year ahead default predictions
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
AUC:76.8%


ROC curve for 5 year ahead default predictions
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
AUC:76.5%