Challenge

Spring 2018 Sports Marketing Analytics Contest: Season ticket churn

Challenge Description

Consider a team that annually plays a 5-game home schedule. Their desire is to be able to identify the fans that are most likely to defect, i.e. not renew their season tickets; so that they can intervene in an attempt to retain those fans. Imagine that they have hired your team to produce a predictive model to help identify these at risk fans.

Your challenge is to build a scoring model that corresponds to the likelihood of fans not renewing their season tickets, based on the 17 variables the team has provided you for customers in their database. Each of the variables serves as a potential predictor variable in the scoring model; or can be used to create new variables, transformed variables, etc. A variety of methodological approaches can be taken. You can build your model totally based on heuristics (see the example in the introduction). Hopefully, you will choose to use a more analytical, data driven approach. Common approaches for problems of this type include a linear probability model, a binary logit model, or a data mining application such as a decision tree. Ultimately, any approach that produces a score for each customer, where higher scores indicate a greater likelihood of churn, can be considered.

Specifically, season ticket holder data was provided for what will be called “Year 1”, which will be used to develop a model to predict defection in “Year 2”. The renewal data for Year 2 will be part of the dataset - this is the calibration data. The validation dataset will include season ticket holder data from Year 2. The scoring scheme developed using the calibration data will then be applied to score the customers in the validation sample, and ultimately predict their defection or renewal in “Year 3”. Year 3 renewal data will not be included in the dataset – this is what you are trying to predict. Once your predictions are submitted, the effectiveness of the Year 3 predictions will be assessed using the actual Year 3 renewal data, as described below.

Evaluation

Each model will be evaluated according to the area under the curve (AUC), a common measure of binary classification model effectiveness. The two components needed to calculate AUC (using the trapezoid method) are sensitivity (the ability of the model to correctly identify defectors) and specificity (the ability of the model to correctly identify those that will renew).

Confusion Matrix

Sensitivity and specificity can be easily derived from a confusion matrix. A confusion matrix counts the number of individuals in each of 4 cells based on the combination of the condition, actual 0s (a renewal) or 1s (a defection); coupled with the prediction, predicted 0s or 1s. This is indicated in the following matrix, along with some common labels for each cell.

	predicted 0 **(renew)**	**predicted 1** **(defect)**
actual 0 **(renew)**	true negatives (tn) “# of predicted 0s that are actually 0s”	false positives (fp) (Type I error) “# of predicted 1s that are actually 0s”
actual 1 **(defect)**	false negatives (fn) (Type II error) “# of predicted 0s that are actually 1s”	true positives (tp) “# of predicted 1s that are actually 1s”

Sensitivity focuses on how well the model identifies individuals that are actually 1s, i.e. defectors. It is also called the “true positive rate (TPR)”, which using the notation in the matrix is TPR = tp/(tp+fn); i.e. the percentage of actual 1s predicted as 1s by the model.

Specificity is similar in nature, but addresses the model’s ability to identify 0s, i.e. those that renew. It is also referred to as the “true negative rate (TNR)”, calculated as TNR = tn/(tn+fp); i.e. the percentage of actual 0s predicted as 0s by the model.

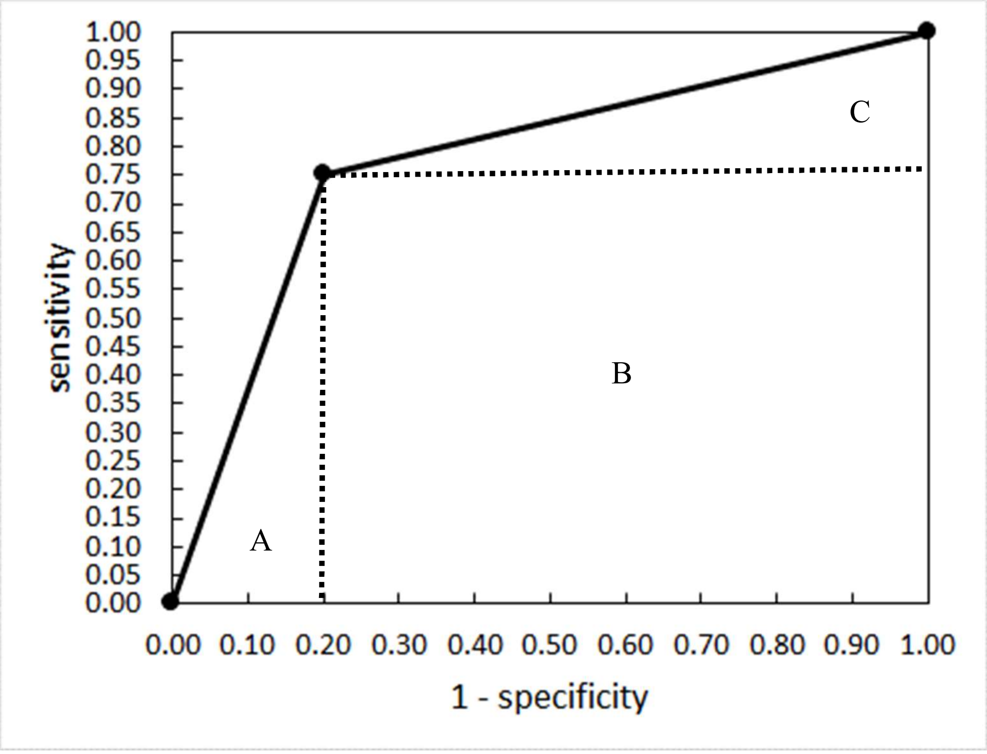
AUC can be measured by simultaneously considering both the sensitivity and the specificity of the model. This is best illustrated graphically.

First consider the following example of a confusion matrix. Please note that given the scores for each individual, the chosen threshold level determines how many predictions of 1s occur, i.e. the confusion matrix cell values are conditional on the chosen threshold level. Your task is to produce the best possible scoring system, and then apply a threshold value that optimizes AUC.

	predicted 0 **(renew)**	**predicted 1** **(defect)**	**total**
actual 0 **(renew)**	true negatives (tn) 33,600	false positives (fp) (Type I error) 8,400	42,000
actual 1 **(defect)**	false negatives (fn) (Type II error) 1,800	true positives (tp) 5,400	7,200
total	35,400	13,800	49,200

The sensitivity (true positive rate TPR), therefore, is TPR = tp/(tp+fn) = 5,400/7,200 = 0.75. The specificity (true negative rate TNR) is TNR = tn/(tn+fp) = 33,600/42,000 = 0.80.

The following graph shows the three areas (A, B, and C) to be summed to produce the area under the curve (AUC) using the trapezoid method. The x-axis is 1-specificity, and the y-axis is sensitivity.



The curve consists of just three points on an x/y grid for the purposes of this contest: (0,0), (1-specificity, sensitivity), and (1,1). The focal point is determined by the threshold level chosen to classify individuals as either 0 or 1. The further to the left the focal point in the graph appears, the better the model is at predicting renewal, i.e. greater specificity. The higher the point appears on the graph, the better the model is at predicting defection, i.e. greater sensitivity. AUC represents the area under this curve. The greater the area, the better the model. A perfect model, with both a specificity and sensitivity of 1, would place the focal point at (0,1) on the grid, and the AUC would attain its maximum value of 1.

The trapezoid method measures the AUC by adding the areas of two triangles and one rectangle, as shown in the graph above. Mathematically, this sum is

$$
\begin{aligned}
\text{AUC} &= \text{area A} + \text{area B} + \text{area C} \\
&= (1\text{-specificity})(\text{sensitivity})/2 + (\text{specificity})(\text{sensitivity}) + (\text{specificity})(1\text{-sensitivity})/2.
\end{aligned}
$$