**Gary Houk for**

**Freedom Financial Network**

Data Science Homework

Background

Some leads are more promising than others. Our sales team would like to rank leads based on the probability that they will “close” successfully, i.e., convert to funded loans. A complicating factor is that close rates typically decline as leads age. The goal of this exercise is to develop a suitable scoring model and then use it to answer some questions.

See code at <https://github.com/TopQuirk67/LeadScoring.git>

Questions (answers replicated with here from the notebook with some selected information; see notebook for complete work)

|  |
| --- |
| 1. **How would you measure the business impact of your model?**   - Outperformance over the relevant metric. If you already have a heuristic for lead scoring, the best way to measure this is an A/B test where some leads are scored by heuristic and others by the model, and measure the conversion rate in each sample  - Additional revenue. You should be able to calculate this from the lift above and the additional revenue you get from it.  - Cost. If the model is very expensive to train and deploy, it may not be worth it for the lift you measure above. On the other hand, it may also be a cost reducer if it can reduce other related costs like staffing, etc.  - Surprise. I like to call them the pork-belly traders, but every firm has the folks who can use their gut to figure out what sales are going to work and which won't. The job of the machine learning algorithm is partially to confirm what these folks already know, but the best result is if you can suprise these domain experts with recommendations that actually lead to new sales. This may be difficult to measure in anything other than an anecdotal way or by polling sales people. |
| 1. **Describe how you would use your model in a business setting.**   Assuming that there is already a method for selecting leads, the initial use case for this algorithm would be as an adjunct recommendation. This would allow the folks who are already deciding which leads to call to have an additional recommendation. If the algorithm is performing much better than historical averages, the next step would be to establish either a team to implement its recommendations or perhaps eliminate the historical method of choosing leads. |
| 1. **Plot the predicted close rate as a function of lead age.**   Chart, bar chart  Description automatically generated We have two non-funded outliers at 60+ days; let's just look at the range up to about 30 days. Let's plot what was asked for, the predicted close rate (i.e. sum of prediction / number in a given bin), but we can also plot the actual close rate as a function of lead age.  Here's how to interpret this plot. On the horizontal axis, we have lead age, and on the vertical axis we plot in blue the average prediction of our model; in green is the average of actual conversion.  We note that the model overpredicts this value compared to ground truth. This is because of class imbalance. We can see this in the confusion matrix above:   * [TN = 23072 FP = 6073] * [ FN = 886 TP = 1299]   So the total numerator for the predbinary prediction in the plot below is FP+TP (how many positive predictions the model makes) while the total numerator for the prediction is FN+TP (how many actual positives there are in the data). The former is larger because the FP's is larger than the FN's, as one might expect for a model struggling with large class imbalance.  So take the scale of the prediction with a grain of salt. |
| 1. **What problems do you foresee when retraining the model a year from now? (We’re interested in biases that might creep into the data.)**   If we implement using the model to select leads, your data will then be biased, presuming that you can't get to every lead. The reason that there will be a bias is that you will be selecting leads the highest scoring leads, and then the conversions you do get to feed back into the model will only consist of leads from this subsample. You will not be determining the conversion from leads that are not selected (presumably low-scoring leads). The only real way to address this is to have a randomized (unbiased) selection group for the retraining dataset. This may not be feasible. |
| 1. **Extra credit: How might you improve upon the greedy method of ranking leads based on predicted close rate? (We have an idea but we don’t know if it’s practical!)**   First, let's check that score actually means something. We will bin the test set by score and make a plot of the actual ("funded") fraction as a function of score.  Chart, bar chart, histogram  Description automatically generated  So, yes, as score increases, so does the yield. If we want to improve the greedy method, it depends on what we mean by "improve." The plot above shows that you will simply get better yield by starting with high-scoring leads, so if yield (at least in the short term) is your goal, it will be hard to improve. There certainly is uncertainty in the score, so one could devise a strategy for more diversity in selection, weighted more to the top scores and decreasingly through the decreasing scores. This would have the added benefit of reducing the selection bias above, so it could improve the process in that way (although honestly I'd have to think pretty hard about how to do the calculation!). But it would necessarily reduce the yield and short-term expense of acquisition. |