Ejercicio I: Churn reduction!

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Scenario and Proposal

You've landed a great job with Green Giant Consulting (GGC), managing an analytical team that is just building up its data science skill set. GGC is proposing a data science project with TelCo, the nation's second-largest provider of wireless communication services, to help address their problem of customer churn. Your team of analysts has produced the following proposal, and you are reviewing it prior to presenting the proposed plan to TelCo. Do you find any flaws with the plan? Do you have any suggestions for how to improve it?

Churn Reduction via Targeted Incentives — A GGC Proposal

We propose that TelCo test its ability to control its customer churn via an analysis of churn prediction. The key idea is that TelCo can use data on customer behavior to predict when customers will leave, and then can target these customers with special incentives to remain with TelCo. We propose the following modeling problem, which can be carried out using data already in TelCo's possession.

We will model the probability that a customer will (or will not) leave within 90 days of contract expiration, with the understanding that there is a separate problem of retaining customers who are continuing their service month-to-month, long after contract expiration. We believe that predicting churn in this 90-day window is an appropriate starting point, and the lessons learned may apply to other churn-prediction cases as well. The

Ejercicio I: Churn reduction!

model will be built on a database of historical cases of customers who have left the company. Churn probability will be predicted based on data 45 days prior to contract expiration, in order for TelCo to have sufficient lead time to affect customer behavior with an incentive offer. We will model churn probability by building an ensemble of trees (random forest) model, which is known to have high accuracy for a wide variety of estimation problems.

We estimate that we will be able to identify 70% of the customers who will leave within the 90-day time window. We will verify this by running the model on the database to verify that indeed the model can reach this level of accuracy. Through interactions with TelCo stakeholders, we understand that it is very important that the V.P. of Customer Retention sign off on any new customer retention procedures, and she has indicated that she will base her decision on her own assessment that the procedure used for identifying customers makes sense and on the opinions about the procedure from selected firm experts in customer retention. Therefore, we will give the V.P. and the experts access to the model, so that they can verify that it will operate effectively and appropriately. We propose that every week, the model be run to estimate the probabilities of churn of the customers whose contracts expire in 45 days (give or take a week). The customers will be ranked based on these probabilities, and the top N will be selected to receive the current incentive, with N based on the cost of the incentive and the weekly retention budget.

Ejercicio II: Whiz-Bang Widget

Your company has an installed user base of 900,000 current users of your Whiz-bang® widget. You now have developed Whiz-bang® 2.0, which has substantially lower operating costs than the original. Ideally, you would like to convert ("migrate") your entire user base over to version 2.0; however, using 2.0 requires that users master the new interface, and there is a serious risk that in attempting to do so, the customers will become frustrated and not convert, become less satistified with the company, or in the worst case, switch to your competitor's popular Boppo® widget. Marketing has designed a brand-new migration incentive plan, which will cost $250 per selected customer. There is no guarantee that a customer will choose to migrate even if she takes this incentive.

An external firm, Big Red Consulting, is proposing a plan to target customers carefully for Whiz-bang® 2.0, and given your demonstrated fluency with the fundamentals of data science, you are called in to help assess Big Red's proposal. Do Big Red's choices seem correct?

Ejercicio II: Whiz-Bang Widget

Targeted Whiz-bang Customer Migration—prepared by Big Red Consulting, Inc.

We will develop a predictive model using modern data-mining technology. As discussed in our last meeting, we assume a budget of $5,000,000 for this phase of customer migration; adjusting the plan for other budgets is straightforward. Thus we can target 20,000 customers under this budget. Here is how we will select those customers:

We will use data to build a model of whether or not a customer will migrate given the incentive. The dataset will comprise a set of attributes of customers, such as the number and type of prior customer service interactions, level of usage of the widget, location of the customer, estimated technical sophistication, tenure with the firm, and other loyalty indicators, such as number of other firm products and services in use. The target will be whether or not the customer will migrate to the new widget if he/she is given the incentive. Using these data, we will build a linear regression to estimate the target variable. The model will be evaluated based on its accuracy on these data; in particular, we want to ensure that the accuracy is substantially greater than if we targeted randomly.

To use the model: for each customer we will apply the regression model to estimate the target variable. If the estimate is greater than 0.5, we will predict that the customer will migrate; otherwise, we will say the customer will not migrate. We then will select at random 20,000 customers from those predicted to migrate, and these 20,000 will be the recommended targets.

Ejercicio I: Solución

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Flaws in the GGC Proposal

We can use our understanding of the fundamental principles and other basic concepts of data science to identify flaws in the proposal. Appendix A provides a starting "guide" for reviewing such proposals, with some of the main questions to ask. However, this book as a whole really can be seen as a proposal review guide. Here are some of the most egregious flaws in Green Giant's proposal:

1. The proposal currently only mentions modeling based on "customers who have left the company." For training (and testing) we will also want to have customers who did *not* leave the company, in order for the modeling to find discriminative information. (Chapter 2, Chapter 3, Chapter 4, Chapter 7)
2. Why rank customers by the highest probability of churn? Why not rank them by expected loss, using a standard expected value computation? (Chapter 7, Chapter 11)
3. Even better, should we not try to model those customers who are most likely to be influenced (positively) by the incentive? (Chapter 11, Chapter 12)
4. If we're going to proceed as in (3), we have the problem of not having the training data we need. We'll have to invest in obtaining training data. (Chapter 3, Chapter 11)

Note that the current proposal may well be just a first step toward the business goal, but this would need to be spelled out explicitly: *see if we can estimate the probabilities well.* If we can, then it makes sense to proceed. If not, we may need to rethink our investment in this project.

Ejercicio I: Solución

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

5. The proposal says nothing about assessing *generalization* performance (i.e., doing a holdout evaluation). It sounds like they are going to test on the training set ("… running the model on the database…"). (Chapter 5)
6. The proposal does not define (nor even mention) what attributes are going to be used! Is this just an omission? Is this because the team hasn't even thought about it? What is the plan? (Chapter 2, Chapter 3)
7. How does the team estimate that the model will be able to identify 70% of the customers who will leave? There is no mention that any pilot study already has been conducted, nor learning curves having been produced on data samples, nor any other support for this claim. It seems like a guess. (Chapter 2, Chapter 5, Chapter 7)
8. Furthermore, without discussing the error rate or the notion of false positives and false negatives, it's not clear what "identify 70% of the customers who will leave" really means. If I say nothing about the false-positive rate, I can identify 100% of them simply by saying everyone will leave. So talking about true-positive rate only makes sense if you also talk about false-positive rate. (Chapter 7, Chapter 8)
9. Why choose one particular model? With modern toolkits, we can easily compare various models on the same data. (Chapter 4, Chapter 7, Chapter 8)
10. The V.P. of Customer Retention must sign off on the procedure, and has indicated that she will examine the procedure to see if it makes sense (domain knowledge validation). However, ensembles of trees are black-box models. The proposal says nothing about how she is going to understand how the procedure is making its decisions. Given her desire, it would be better to sacrifice some accuracy to build a more comprehensible model. Once she is "on board" it may be possible to use less-comprehensible techniques to achieve higher accuracies. (Chapter 3, Chapter 7, Chapter 12)

Ejercicio II: Solución

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Flaws in the Big Red Proposal

We can use our understanding of the fundamental principles and other basic concepts of data science to identify flaws in the proposal. Appendix A provides a starting guide for reviewing such proposals, with some of the main questions to ask. However, this book as a whole really can be seen as a proposal review guide. Here are some of the most egregious flaws in Big Data's proposal:

Business Understanding

- The target variable definition is imprecise. For example, over what time period must the migration occur? (Chapter 3)
- The formulation of the data mining problem could be better-aligned with the business problem. For example, what if certain customers (or everyone) were likely to migrate anyway (without the incentive)? Then we would be wasting the cost of the incentive in targeting them. (Chapter 2, Chapter 11)

Data Understanding/Data Preparation

- There aren't any labeled training data! This is a brand-new incentive. We should invest some of our budget in obtaining labels for some examples. This can be done by targeting a (randomly) selected subset of customers with the incentive. One also might propose a more sophisticated approach (Chapter 2, Chapter 3, Chapter 11).
- If we are worried about wasting the incentive on customers who are likely to migrate without it, we also should observe a "control group" over the period where we are obtaining training data. This should be easy, since everyone we don't target to gather labels would be a "control" subject. We can build a separate model for migrate or not given no incentive, and combine the models in an expected value framework. (Chapter 11)

Modeling

- Linear regression is not a good choice for modeling a categorical target variable. Rather one should use a classification method, such as tree induction, logistic regression, k-NN, and so on. Even better, why not try a bunch of methods and evaluate them experimentally to see which performs best? (Chapter 2, Chapter 3, Chapter 4, Chapter 5, Chapter 6, Chapter 7, Chapter 8)

Ejercicio II: Solución

Evaluation

- The evaluation shouldn't be on the training data. Some sort of holdout approach should be used (e.g., cross-validation and/or a staged approach as discussed above). (Chapter 5)
- Is there going to be any domain-knowledge validation of the model? What if it is capturing some weirdness of the data collection process? (Chapter 7, Chapter 11, Chapter 14)

Deployment

- The idea of randomly selecting customers with regression scores greater than 0.5 is not well considered. First, it is not clear that a regression score of 0.5 really corresponds to a probability of migration of 0.5. Second, 0.5 is rather arbitrary in any case. Third, since our model is providing a ranking (e.g., by likelihood of migration, or by expected value if we use the more complex formulation), we should use the ranking to guide our targeting: choose the top-ranked candidates, as the budget will allow. (Chapter 2, Chapter 3, Chapter 7, Chapter 8, Chapter 11)

Of course, this is just one example with a particular set of flaws. A different set of concepts may need to be brought to bear for a different proposal that is flawed in other ways.

Data Science y la Estrategia de Negocio

Eduardo M. Sánchez Vila

eduardo.sanchez.vila@usc.es

CITIUS

Grupo de Sistemas Inteligentes

Universidad de Santiago de Compostela