**How to retain your customers - churn prediction in the automotive service business**

Churn prediction with help of supervised machine learning modelling is becoming ever more common, and really, it's no wizardry. But prediction is only the first step towards retention. And so, every churn project should address three questions:

* Can we effectively predict churn?
* What are the most important features leading to churn?
* How should we act?

After a very short introduction to the data and model, I will do exactly that based on a project in the automotive service business.

Short introduction to the project

The goal is to identify and proactively address customers with high risk for churn. The underlying dataset contains historical data of about 50'000 cars that have been sold by a large European automotive retail group and then came back for aftersales service at least once.

All instances have been classified as either 'ACTIVE' (active customers at the time the dataset was created) and 'CHURN' (customers that have not turned up for at least 27 months after their last visit at the time the dataset was created.) It is to note that the classes are imbalanced, 32% of the instances are labeled as CHURN, 68% as ACTIVE.

For making the predictions 42 input variables are present, that can roughly be grouped as follows:

* Features of a car's history like age, mileage and number of past service events
* Features of technical car specifications like make, model, number of gears and so on
* Features of the car's owner like age and some socio-demographic data
* Features of the company branch the car is affiliated with, including the distance to it's home location

A GradientBoostingClassifier from Python's scikit-learn library was chosen for prediction modelling. As expected it dealt very well with outliers, the skewedness of numerical data and the many categorical variables. (For more details you can find the complete source code on GitHub.)

Communicate the results

Question 1: Can we predict churn?

Yes, we can. "Despite of an underrepresented 'CHURN' class and low feature correlation in regard to the target variable our model achieves an F-beta-0.5-score of 0.7307 and the AUC of the ROC is an impressive 0.881 …" Uh, great. But not very intuitive. (I mean, if you're not used to it, even a confusion matrix can be quite confusing.)

If I have to explain model performance to a non-math-savy audience, I try to refer to the status quo as benchmark: " If today we randomly address 100 customers as potential churners, we will have a hit rate of only 32% (due to the distribution of our target variables). But: If we address 100 potential churners based on the predictions of our model, then we achieve an average hit rate of 76% (the so called 'precision'), which is more than twice as high as today."

But of course we also have to admit that the model identifies only 64% of all effective CHURN instances as such on average(the so called 'recall'). To communicate the results and the important relationship between precision and recall I personally prefer the Precision-Recall-Curve to the more famous but harder to interpret ROC-Curve.

[Fig-1: Precision-Recall curve, showing the trade-off between recall and precision.]

Question 2: What are the most important features leading to churn?

This is an extremely important question concerning the business and hiding behind the 'black box AI' is not an option. Many algorithms allow for some interpretation of the feature weights.

Tree-based models like the GradientBoostingClassifier are especially suited to this (see below for the results in this project). It is always advantageous to work with at least one interpretable algorithm. But even if that does not work, one can try to create an intuition about the most important correlations and connections through a PCA or at least the initial EDA. Doing so will help you on the next question.

[Fig-2: Feature importances]

That's were business value is created. A good starting point is to analyze the most important features in more detail and to relate them to each other. In our case for example, it was surprising to see that two variables (vehicle age and duration of the customer relationship) that are more than 90% correlated, account for roughly 60% of the model's cumulative feature importance.

Question 3: How should we act?

On closer examination, I found that subtracting one from the other gives an estimate of the last visit's recency. After consultation with the marketing department we concluded that customers with vehicles older than 2 years who had no service appointment in 200 days have an extremely high risk for churn. Customers who are about to meet those criteria are now flagged with some advance (100 days after the last service visit), evaluated with help of our model and depending on the predicted risk for churn they are approached with a costlier or less costly retention measure.

Take-away

Predictive modeling should not only aim to shine with the highest possible metrics. It is as important to interpret the models correctly and to derive concrete applicability in everyday business processes. Doing so will also facilitate communication with your non-tech-savvy colleagues and strengthen the trust in what you do.

Honestly, what is the use of a top-performing model that can predict with certainty that a customer will migrate to the competition? He'll be gone. You better talk to him before it's too late – and prove your model wrong by keeping him loyal.

(This article was written as part of a task in Udacity's Data Science Nanodegree.)