

Customer Churn Prediction

A telecommunications case study

By: Pablo González

Why customer churn?

- Relevant and challenging problem affecting even the largest, most stable companies.
- Crucial for maintaining a sustainable business growth.
- The costs of acquiring new customers can be higher than the ones of retaining loyal ones.



An illustration on a solid blue background. On the left, a yellow ladder stands vertically. On the right, a man with dark hair, wearing a white long-sleeved shirt and dark blue trousers, stands on a small white rectangular pedestal. He is holding a black pen in his right hand and pointing it towards a white line graph. The graph starts at the top left and trends downwards with several small peaks and valleys, ending with an arrowhead pointing towards the man. The text is centered in the middle of the image.

"It is 6-7 times more expensive to acquire a new customer than it is to keep a current one"

Statement attributed to the White House Office of consumer affairs

Business Problem

How to better predict which customer is likely to churn?

In other words, we want to know **who is likely** to turnover and be able to target him with a specific campaign.

Performance measures: Testing error, AUC and Sensitivity.

VALUE

Data Source

- Telco Customer Churn data set.
- 7043 observations and 21 variables
- Contains information about customer churn, contracted services and demographic characteristics, among others.

kaggle

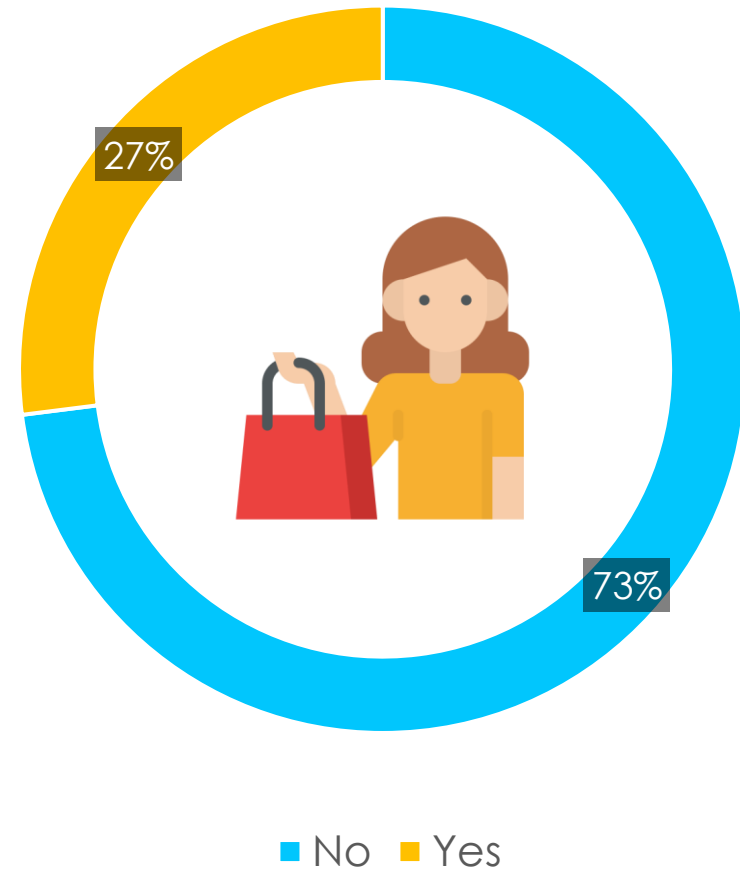


kaggle.com/blastchar/telco-customer-churn

Preliminar Analysis

- 11 missing values
- No outliers (IQR method)
- Response variable: Churn

Customer Churn



Potential predictors

Ranked Cross-Correlations

10 most relevant

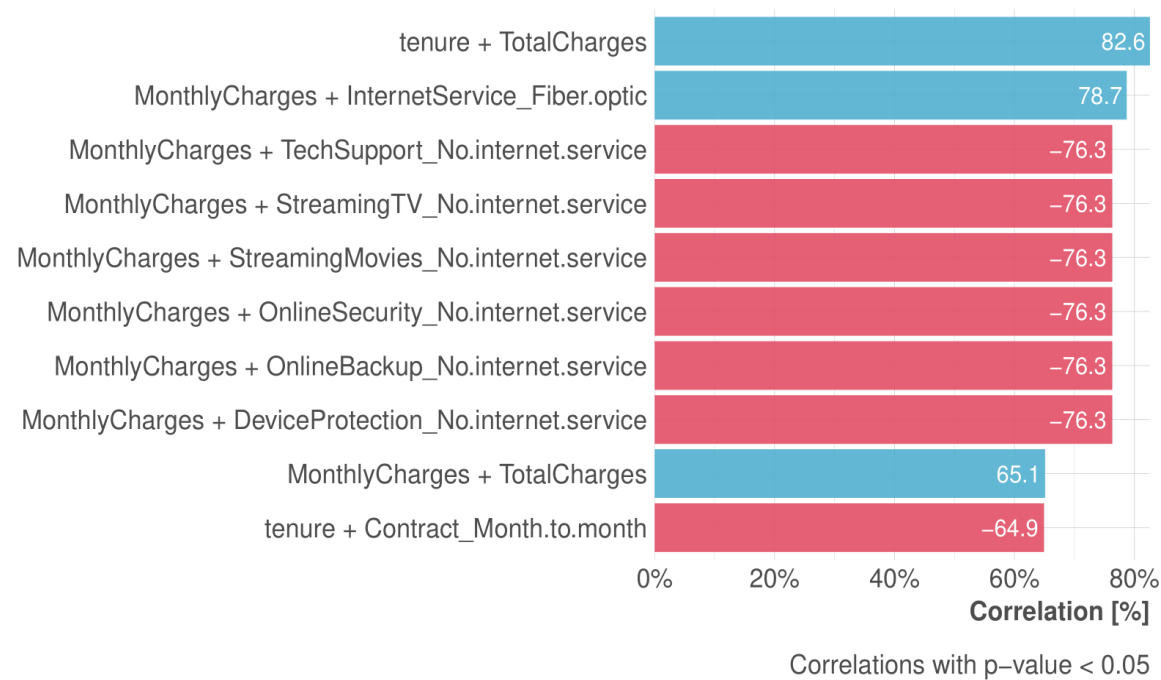
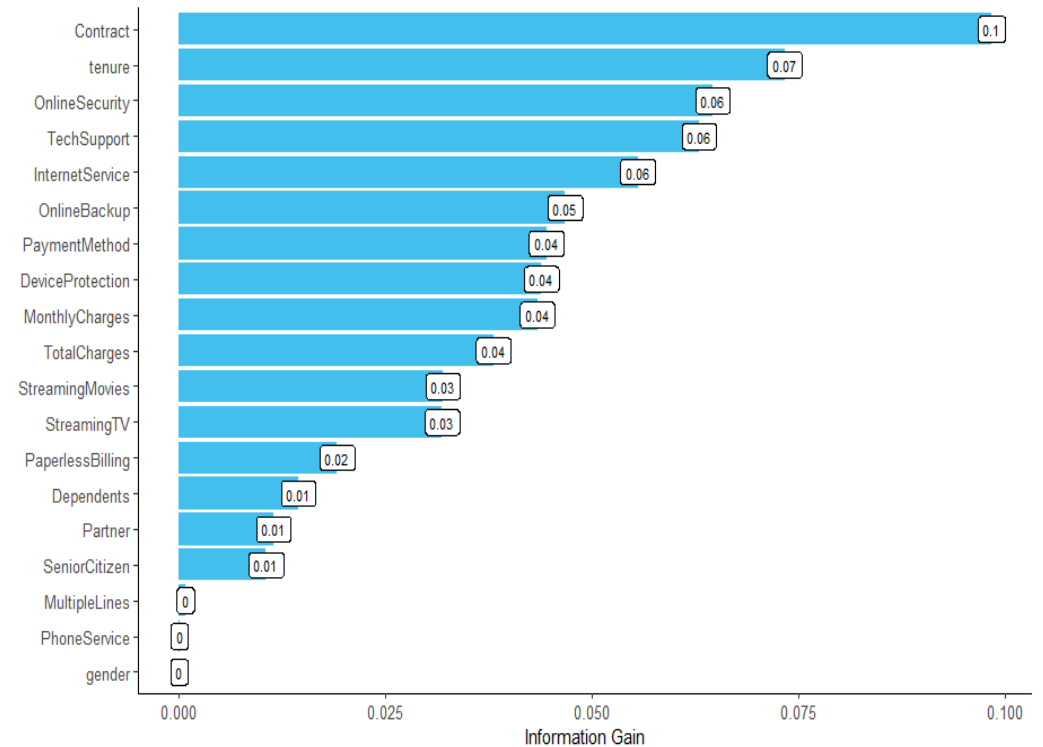


Figure 1: Predictors by information gain



Methodology

Used models:

- Naive Bayes Classifier
- Logistic Regression
- Random Forest
- Boosting
- Support Vector Machines

Important clarifications:

- Train-test Split: 9:1
- 100-fold MC cross-validation
- Testing error, AUC, Sensitivity
- T-test and Wilcoxon-test



Results: Misclassification Error (Testing)

	Mean	Median	Variance
Naive Bayes	0.27185	0.27454	0.00025
Logistic Regression	0.19502	0.19488	0.00018
Random Forest	0.24371	0.24395	0.00028
Boosting	0.19565	0.19488	0.00020
Support Vector Machines	0.31664	0.31579	0.00030

Results: Area under the ROC curve (AUC)

	Mean	Median	Variance
Naive Bayes	0.81949	0.81973	0.00030
Logistic Regression	0.84655	0.84804	0.00027
Random Forest	0.84366	0.84273	0.00029
Boosting	0.84961	0.85067	0.00028
Support Vector Machines	0.82758	0.83539	0.00498

Results: Sensitivity

	Mean	Median	Variance
Naive Bayes	0.80763	0.80878	0.00080
Logistic Regression	0.55223	0.55076	0.00137
Random Forest	0.78972	0.78847	0.00084
Boosting	0.58436	0.58220	0.00155
Support Vector Machines	0.86230	0.86163	0.00063

Conclusions

- Naive Bayes, Random Forest and Support Vector Machines are the best performing models overall.
- Among them, the Random Forest classifier provides a significantly better performance in terms of testing error and AUC.
- The Support Vector Machines classifier is the best if we only care about sensitivity.

The background is a light blue gradient. In the center-left, a blue silhouette of a businessman in a suit stands on a large red paper airplane, holding a telescope to his eye. To the right, a white line graph with red circular markers trends upwards, ending in a large red arrowhead. The background is decorated with several white, stylized clouds, some of which are hanging from thin white lines. Faint blue vertical bars, resembling a bar chart, are visible in the background.

THANK YOU FOR YOUR ATTENTION!

Image Sources

- www.qminder.com/customer-retention-strategies/
- www.retently.com/blog/three-leading-causes-churn/
- www.credilink.com.br/as-vantagens-do-churn-rate-para-seu-negocio/
- www.indyme.com/wp-content/uploads/2020/11/customer-icon.png
- www.vectorvest.ca/number-1-rule-to-investing-let-the-trend-be-your-friend/