Telco Customer Churn: Predictive Modeling and Explainability

Final Project – XAI

Antonio Lorenzo Díaz-Meco

December 4, 2025

1 Introduction

This project tries to address a classic problem that companies in the telecomunicatios sector face: predicting which customers are more likely to get out of the service (churn). For a telecom company, losing a client doesn't only mean to lose some monthly income, it is also a cost for the future, because it is usually way more expensive to find new customers than to keep the existing ones.

The interest is not only trying to predict who wants to leave, it is also important to understand why would they leave. In this case, explainability is as relevant as accuracy. Areas like business, marketing and retention need to have clear explanations to be able to justify decisions, know where to put the resources and design effective campaigns. Also, different regulations make it mandatory that models that affect end users are interpretable.

The main goal of this project is to train a model that can robustly predict churn and also have some XAI techniques that help us to understand how the model works globally and also how it makes individual predictions. Two sanity checks have also been implemented to check if the explanations really show the model behaviour or if they are based on artifacts.

Stakeholders and Supported Decisions

This model was thought mainly for three profiles of those companies:

- **Retention team**: they need a list that priorizes clients with high churn risk to decide who to call, what to offer and when to contact them.
- **Marketing team**: uses global patterns (contract type, tenure, internet service) to design campaigns and specific packages by segment.
- **Management and Risk**: wants to understand why the churn rate rises or falls in certain groups, and evaluate the economic impact of intervening on them.

In this context, explainability isn't an extra, it is a condition for the models to be used. The marketing team should be able to respond to questions like: *"why is this client marked as high risk?"* or o *"what features could we change to reduce the probability of churn on this exact segment?"*.

2 Data and Methods

2.1 Dataset

The dataset that I used comes from *Telco Customer Churn Dataset*. It has more than 7k clients and a mix of many different features: the type of service, payment method, tenure, monthly and total charges, among others.

The preprocessing included:

- Converting `TotalCharges` to numeric and removing rows with invalid values.
- Separating between numerical and categorical variables.
- Standardization of numerical variables.
- One-hot encoding for categorical variables.

Also, I divided the dataset into train and test sets using stratified sampling to keep the original churn proportion (approx. 26%). It is really disbalanced as we can see, so we will have to deal with it later on our models and metrics.

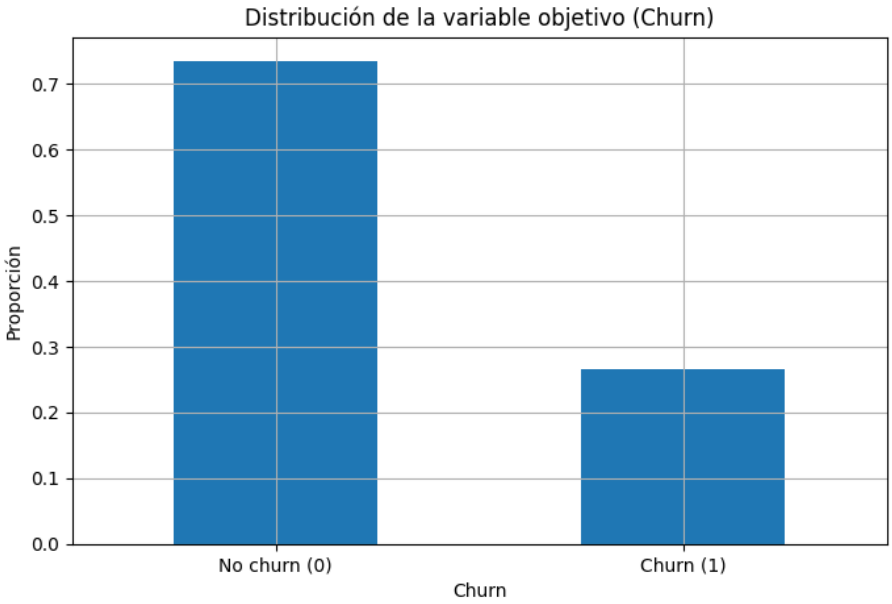

Figure 1: Distribution of the target variable (Churn).

2.2 Models

Two main models were trained:

- **Logistic Regression** with `class_weight="balanced"`.
- **Random Forest** with grid search over the hiperparameters.

The metrics I used are: accuracy, balanced accuracy, ROC-AUC, PR-AUC, accuracy, recall and F1. Particularly, PR-AUC is critical because the churn problem is really unbalanced.

2.3 Explainability Techniques

Three XAI techniques were applied:

1. **Permutation Feature Importance (PFI)**: measures how the performance is affected when permuting columns.
2. **SHAP global**: calculates the mean absolute contribution of each transformed feature.
3. **SHAP local**: it breaks down an individual prediction into its main contributions.

Finally, two sanity checks were made:

- Retraining the model after eliminating key features identified with XAI techniques.
- Retraining the model after shuffling the labels to confirm that there is no artificial signal or data leakage.

3 Results

3.1 Performance of the models

The Table 1 summarizes the main metrics of both models.

Model	Accuracy	BalAcc	ROC–AUC	PR–AUC	Accuracy	Recall
LogReg	0.726	0.748	0.835	0.618	0.490	0.797
RandomForest	0.758	0.749	0.829	0.632	0.533	0.730

Table 1: Comparison of metrics between Logistic Regression and Random Forest.

As we can see in Table 1. The Random Forest improves the PR–AUC and the accuracy on the positive class, while the logistic regression achieves a higher recall. Given that this metrics already reflect good enough the behaviour of each model, the ROC and Precision-Recall curves won't be included here, as they won't give us many additional conclusions. If you want to see them, they are on the notebook.

3.2 Global Explainability

Permutation Importance showed a clear pattern: **Contract** and **InternetService** are the features that affect the performance of the model the most. All the other features have much less impact.

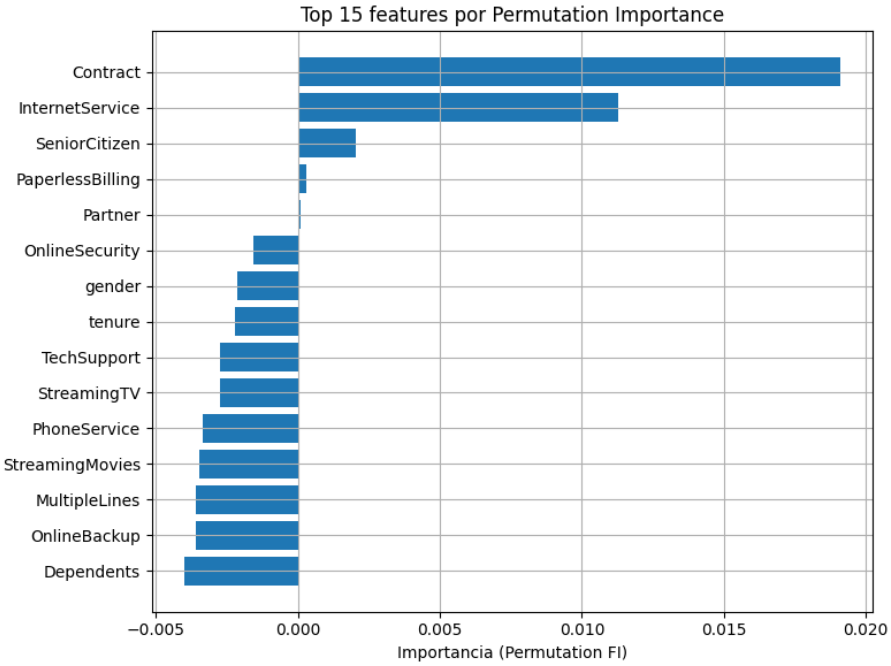

Figure 2: Top features according to Permutation Feature Importance.

The SHAP global values confirmed this pattern, with the dummies of contract and the features related to the tenure and charges taking the first places.

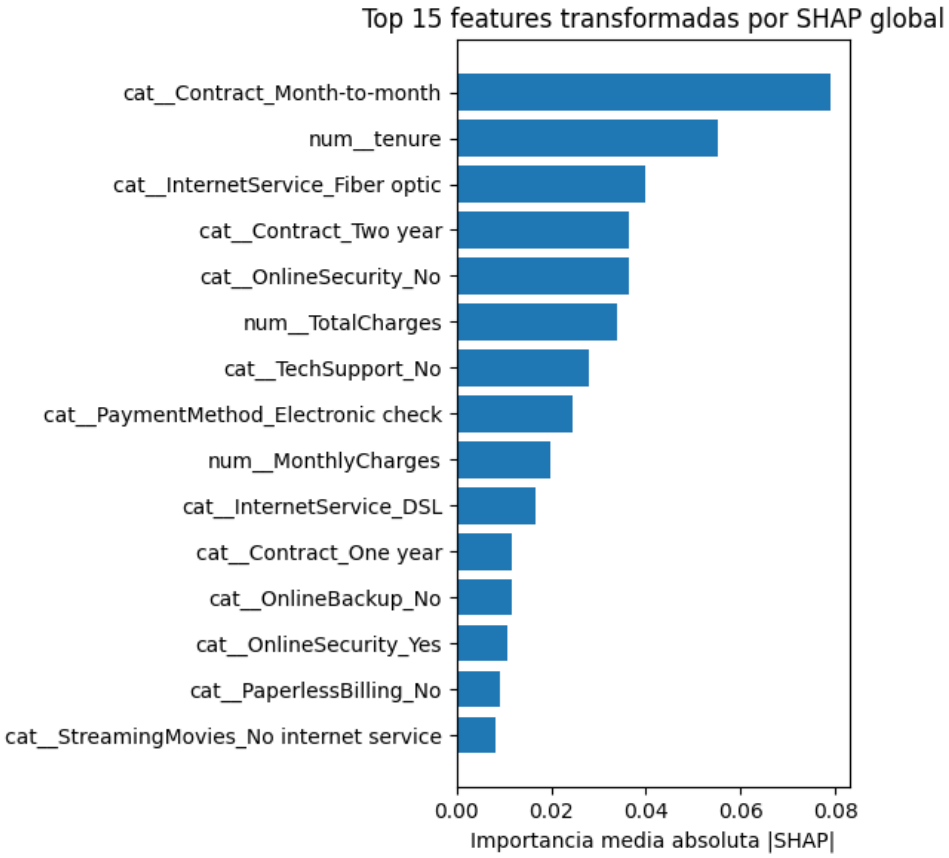

Figure 3: Top features transformed according to global SHAP.

3.3 Local Explainability

The local explanation seen with waterfall plots allows us to justify individual predictions. In clients with high churn risk the most important features normally are monthly contracts, low tenure and some services such as fiber optic.

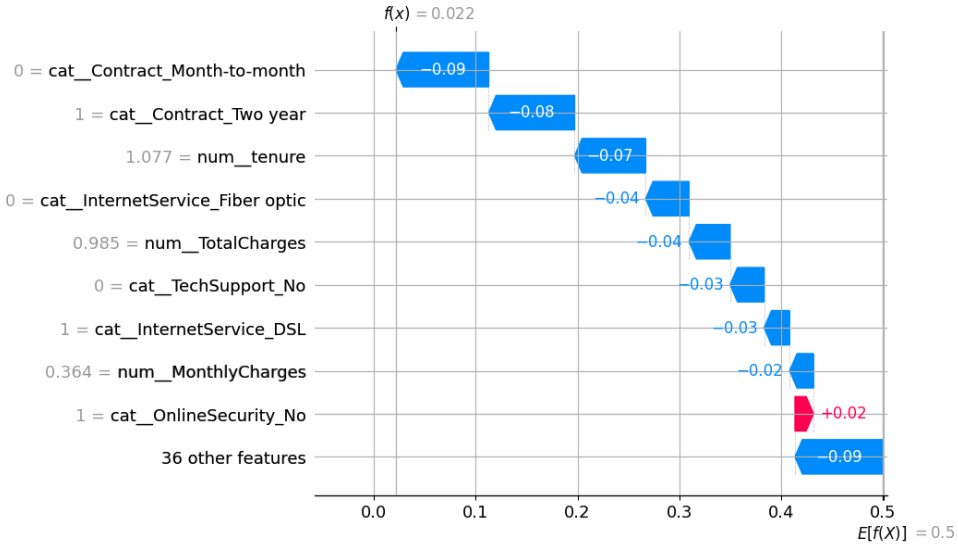

Figure 4: Ejemplo de explicación local con SHAP.

3.4 Sanity checks

The first sanity check demonstrated that when removing key features such as `Contract` or `tenure`, the performance of the model clearly falls, especially in PR-AUC. The second sanity check confirmed that, if we shuffle the labels, the model colapses to a random behaviour, with discards the presence of *data leakage*.

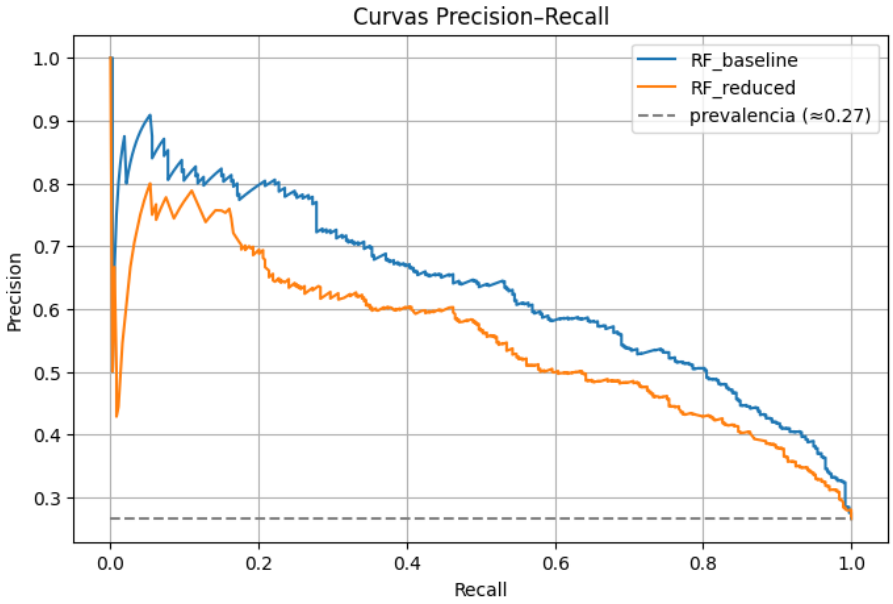

Figure 5: Comparison between RF baseline and RF reduced.

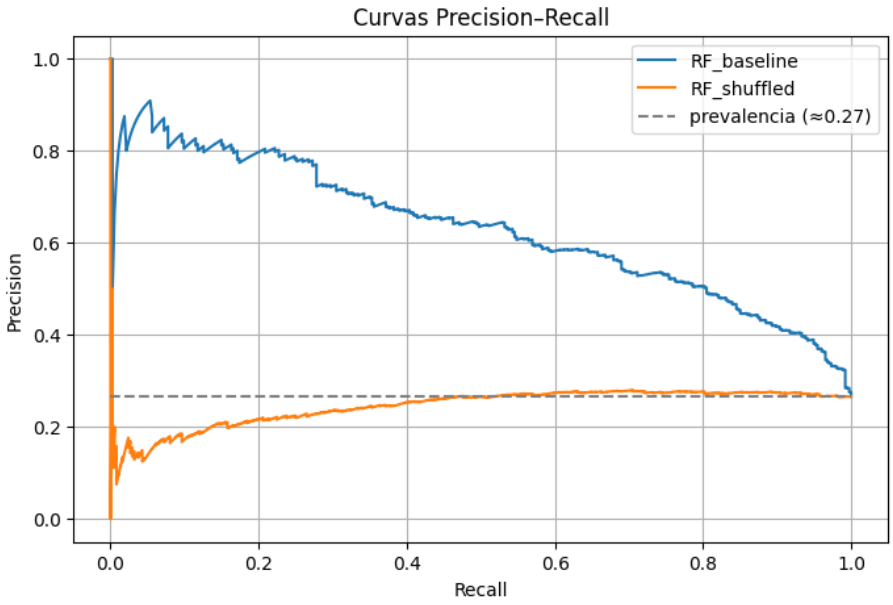

Figure 6: Comparison between RF baseline and RF with suffled labels.

4 Actions and Insights

4.1 Business actions suggested by the explanations

Based on the results from PFI and SHAP, there are several practical actions that the company could take:

- **Month-to-month contracts**: customers on month-to-month plans are clearly the most likely to churn. A reasonable idea would be to offer them discounts for switching to a yearly plan or give them some extra benefits (for example, better customer support). This is especially important for those who also pay high monthly fees.
- **Fiber-optic customers with low tenure**: the combination of `InternetService = Fiber optic` and low `tenure` seems particularly unstable. For this segment, it might be worth checking for service issues, reviewing prices, or even setting up specific follow-up campaigns during the first months.
- **Long-standing but expensive customers**: SHAP shows that high monthly charges increase the probability of churn, even for customers who have been with the company for a long time. In these cases, revising the pricing structure or offering more affordable bundles could help prevent churn.

These ideas don't come just from looking at the raw data; they are supported by the global importance of the features and by the local explanations for individual customers.

4.2 Use of explanations to simplify the model

The first sanity check, in which the model is retrained without some features such as `Contract` or `tenure`, clearly gets worse on PR–AUC and the quality of the curves of Precision-Recall.

But, it also reveals that other features with really low importance almost give no information to the model.

In the real world, this would allow:

- Reducing the number of input features, simplifying the pipeline and manteinance.
- Avoid collecting data that has almost zero effect on the prediction, saving some operative costs.
- Focusing the efforts on monitorizing a reduced subset of key features.

5 Discussion

5.1 Limitations

On top of all the limitations that the explanation methos have, the dataset is also relatively small and highly depends on categorical features. Many of those features are correlated, which could make it more difficult to interpret the model. Also, there is no temporal history available, which would be very valuable in churn.

5.2 Risks

There is always a risk of treating correlations as if they implied causality. It's also important to be careful when acting on specific customer groups to avoid unwanted side effects or unfair biases, especially when sensitive variables such as age or income level are involved.

5.3 Critical reflection on XAI

The explainability techniques used in this project are not neutral or perfect. **Permutation Feature Importance** works at the level of the original columns and can underestimate the relevance of highly correlated variables. When only one of them is permuted, part of the signal is still captured by the remaining correlated features. This happens in our dataset with several Internet-related services and add-ons that tend to appear together. For this reason, it is useful to contrast PFI with **SHAP**, which operates in the one-hot encoded space and captures local interactions in a more detailed way.

Another important point is that these explanations are purely associative. The fact that month-to-month contracts show up as a strong risk factor does not automatically mean that changing a customer's contract type would "fix" churn. A business stakeholder who interprets the plots too literally might design unrealistic or overly aggressive interventions, placing too much confidence in what the model appears to suggest.

The *sanity checks* help calibrate our trust in the explanations. When we remove variables identified as important, the model performance drops exactly in the direction that PFI and SHAP anticipated, which indicates a reasonable level of faithfulness: when we take away what the model relies on, it actually performs worse. When the labels are shuffled, the model collapses toward random behavior and the metrics fall to the baseline prevalence. This rules out data

leakage and shows that the explanations of the baseline model are grounded in real structure rather than noise.

Finally, it is important to keep in mind that some features may touch sensitive areas. The presence of `SeniorCitizen`, for example, suggests that explanations must be used carefully in practice: the goal should be to improve customer experience, not to justify unfair or discriminatory treatment toward specific groups.

5.4 Future Work

There are a few ways this project could be taken further:

- Trying sequence-based models if we ever get historical data, since churn often depends on how a customer behaves over time.
- Looking at fairness to make sure the model isn't affecting certain groups in an unfair way.
- Using counterfactual explanations to give more personalized suggestions for each customer.
- Adding monitoring tools in production to detect *data drift* and make sure the model stays reliable.

6 Conclusion

This project shows that we can build a churn model that works reasonably well and also understand the reasons behind its predictions. Both PFI and SHAP point to the same important features, and the sanity checks suggest that the explanations are actually reflecting how the model behaves and not just noise. Overall, the mix of performance, interpretability, and the checks we ran makes this approach a good fit for a real telecom setting.