Churn Prediction

From dissatisfaction to loyalty

Andrea Pan - 514555

The problem

Context

The company has reported **increasing risks of churning**, negatively impacting the order value and overall customer lifetime value.

Objectives

- **Identify** customers in ‘at-risk’ segment
- **Explain** the reasons behind the growing churn rate
- **Apply** strategies to turn unhappy customers to loyal ones

Challenges

- Imbalanced data
- Interpretability
- Changing behavior
- Actionability

Solution

Machine Learning Model

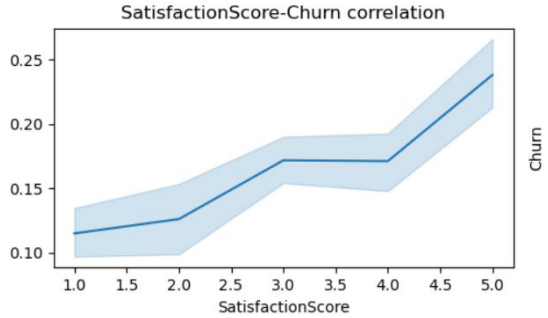
We propose a **machine learning classification model**, capable of labeling customers in risk of churning with high consistency.

This approach is not only accurate, capturing non-linear relationships, but also **explainable**: we can build strategies around the predictors of the model.

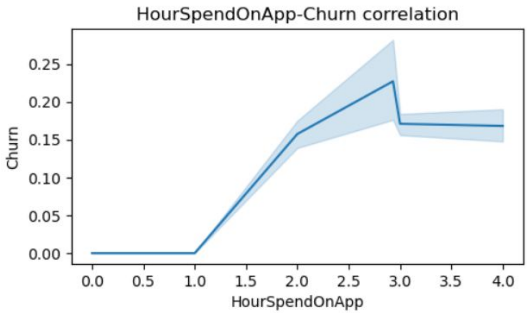
The dataset

Results from Exploratory Data Analysis (EDA)

Feature	Expected Relevancy
Tenure	High
WarehouseToHome	Situational
HourSpentOnApp	High
NumberOfDeviceRegistered	Unknown
SatisfactionScore	High
NumberOfAddress	Unknown
OrderAmountHikeFromlastYear	High
CouponUsed	Situational
OrderCount	High
DaySinceLastOrder	High
CashbackAmount	Situational
PreferredLoginDevice	Low
CityTier	High
PreferredPaymentMode	Low
GenderDistribution	Low
PreferredOrderCat	Unknown
MaritalStatus	Low
Complain	High



Unexpected correlations



5630 entries
1860 total missing values
20 columns

Aliases merged

- Some data relationships are non-linear
- Overall weak linear correlation
- Most significant features:
  1. Tenure
  2. DaySinceLastOrder
  3. Complain

Results from Exploratory Data Analysis (EDA)

We discovered a strong association between Churn and PreferredOrderCat. Thus, the question is:

What categories of order are the most likely to churn? What about their tenure and order count?

It appears that we have many one-timers, with many of them buying a mobile phone or a laptop and then never returning.

Preferred Order Category	Tenure	Order Count	No. of Churn
Mobile	1	2	191
Mobile	0	1	129
Mobile	10	1	68
Laptop & Accessory	1	2	40
Laptop & Accessory	0	1	40
Mobile	1	3	28
Mobile	0	2	23
Laptop & Accessory	0	2	14
Fashion	0	1	14
Fashion	1	2	14

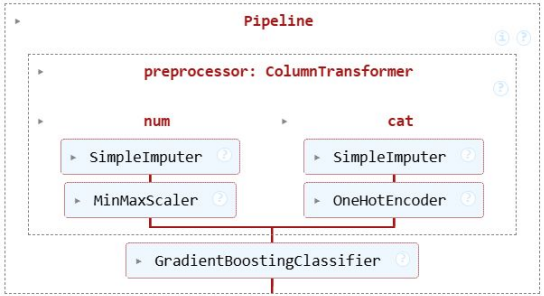
We can already make a suggestion for a strategy to take: Make offers for accessories on their recently bought mobile or laptop to retain engagement.

The model

The Workflow

Pipeline Definition

Structure:



Grid Search

Performed for both **finding the best model**, and for **hyperparameter optimization**.

We tested Random Forest, Linear Regression and Gradient Boosting.

Linear Regression was discarded.

Hyperparameter Tuning

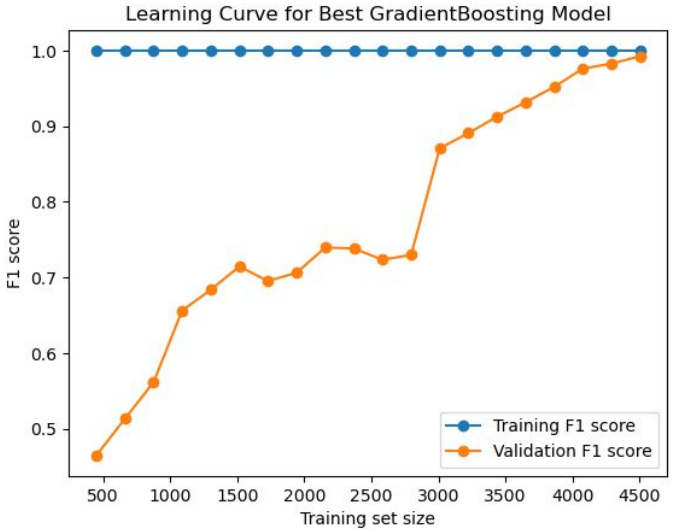
After much optimization of hyperparameters of both Random Forest and Gradient Boosting (a total of 3060+ fits!) we decided on the latter.

Note: **Both performed flawlessly**, but Gradient Boosting was marginally better in training.

Model Results

Gradient Boosting	Precision	Recall	F1 Score	Support
0 (not churn)	1.00	1.00	1.00	1172
1 (churn)	1.00	1.00	1.00	236
Accuracy			1.00	1408
Macro avg	1.00	1.00	1.00	1408
Weighted avg	1.00	1.00	1.00	1408

Confusion Matrix	Predicted Negative	Predicted Positive
Actual Negative	1172	0
Actual Positive	0	236



The classification report is based on the test set: we can be sure the **model is not overfitted**.

To further prove it, we plotted the learning curve, and we can see the validation score keeps raising.

The model's performance jumped dramatically once the critical mass of around 3000 training size.

Any less data, and we'd have a suboptimal model.

Explainability & Strategies

SHAP Explainability

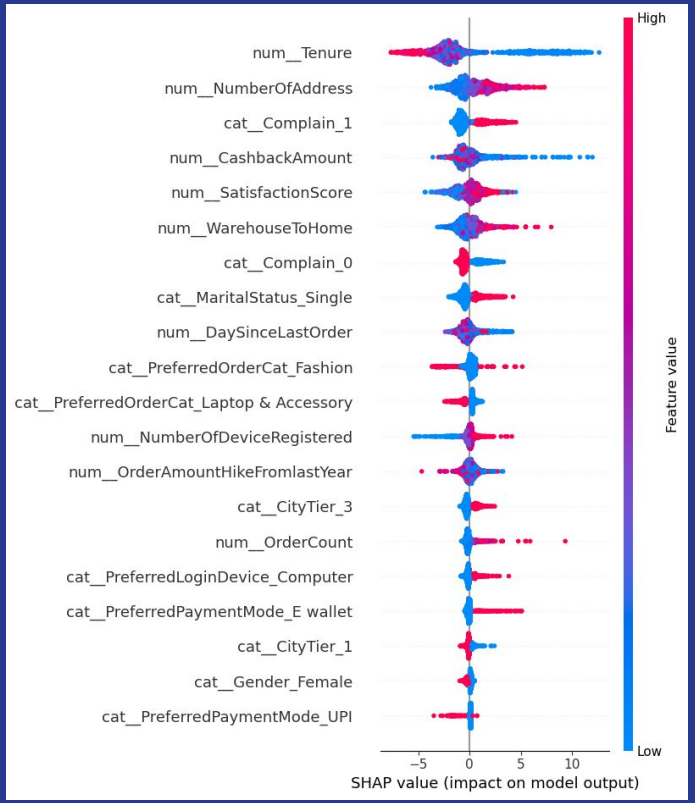
SHAP (SHapley Additive exPlanations) is a method used to explain how machine learning models make predictions.

How it works:

- Each feature is considered a player
- Each player contributes to the prediction
- A payout is divided across players based on contribution

SHAP works really well with Gradient Boosting Models.

SHAP



Suggested Strategies

At-Risk Segment

Low Tenure

Onboarding campaign and loyalty programs.

High Number of Addresses

Might be movers. Offers based on home utilities or free delivery service.

Complaints

Complaint Resolution

Email focusing on empathy, accountability, and give offers to turn dissatisfaction into loyalty.

Personalization

Based on Demography or Geographics

Singles and Tier 3 Cities are more likely to churn in our data.

Research about personal/regional preferences, and give bonuses through referral.

Watch out!

Cashback has most SHAP data points on 0, but there's a blue trail that leans into higher SHAP values.

High cashback: no churn reduction.

Low cashback: high chance of churn.

Consider other strategies.

Thanks for the attention!