

ENSSEA

Data Mining & Machine Learning for Customer Churn Prediction Case: Telecom Company

Author:
YACINE AMMI

Supervisor:
KHALED ROUASKI

Presentation Plan

1. CONCEPTUAL FRAMEWORK

- Introduction
- Objectives of the study
- Research questions and Hypotheses
- Literature Review
- Methodology
- Tools and programs



2. EMPERICAL STUDY

- Dataset overview
- Exploratory data analysis
- RFM clustering
- Preprocessing
- Model selection
- Model validation
- Model evaluation
- Model interpretation
- Conclusion
- Recommendations



1

CONCEPTUAL FRAMEWORK

Introduction

“

Acquiring a new customer is anywhere
from **five** to **25** times more **expensive**
than **retaining** an **existing** one

”

—Amy Gallo, Harvard Business Review

Objectives of the study

01

Build a precise predictive model for churn prediction

02

Determine the primary factors of customer churn

03

Give recommendations to the company to reduce churn



Research questions and Hypotheses



01

- ❑ Can data mining techniques accurately forecast customer behaviour?
- We can rely on data mining to anticipate client behaviour.

02

- ❑ What algorithms are most effective in customer churn prediction?
- Ensemble Learning methods are the most suited algorithms to predict customer churn.

03

- ❑ What is the most crucial variable in predicting the customer churn given the dataset?
- The customer last account balance is the most crucial variable for predicting customer churn,

Literature review

CRM

Is an essential business strategy that seeks to establish and maintain profitable relationships with customers

RFM

Stands for Recency, Frequency and Monetary, it is a three-dimensional method of categorizing or evaluating clients in order to discover the best customers.



Churn

Is a marketing term that refers to a customer who switches from one company to another.

Ensemble Learning

is a technique that combines many simple models to create a single, potentially highly powerful model

Methodology



PHASE 1
Data Exploration



PHASE 3
Data Preprocessing



PHASE 5
Evaluation &
Interpretation



PHASE 2
RFM Clustering



PHASE 4
Classification



PHASE 6
Recommendations

Tools and programs





2

EMPERICAL STUDY

Dataset overview

The dataset we used in this study is a public dataset about **customers** of an anonymous **telecommunications** company, It was collected from **Kaggle**, an open-source platform that hosts datasets and **data science** projects.

- This dataset is composed of:

66,469

Customer

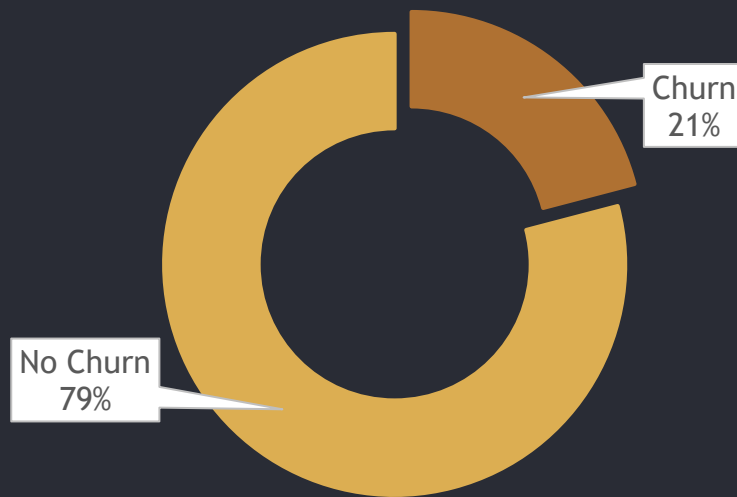
66

Variables

Exploratory Data Analysis

Before starting the study, we did some preliminary steps as follows:

- Dropped the Year and user_id columns
- Dropped the users with no activity for more than 1 year
- Dropped 17 variables that had correlations greater than 85%



Non churned customers:

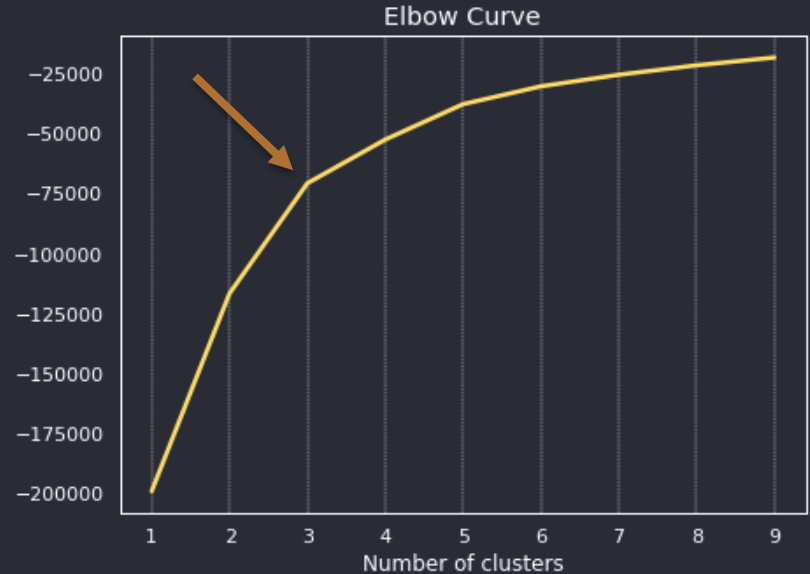
52,562

Churned customers:

13,907

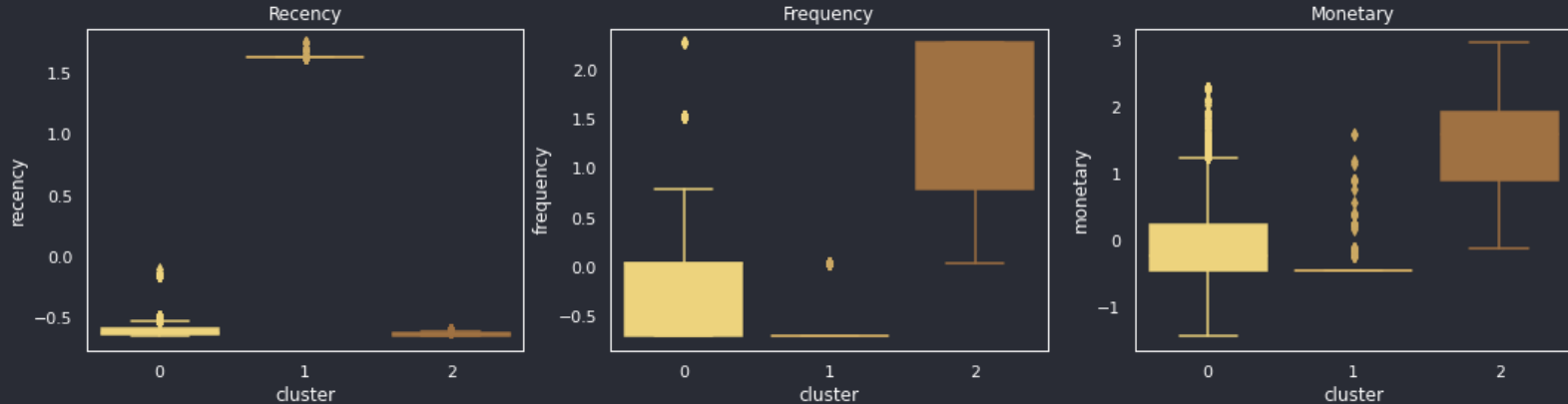
Clustering by RFM

- We note that $k = 3$ is the ideal number of clusters according to the elbow method,



Clustering by RFM

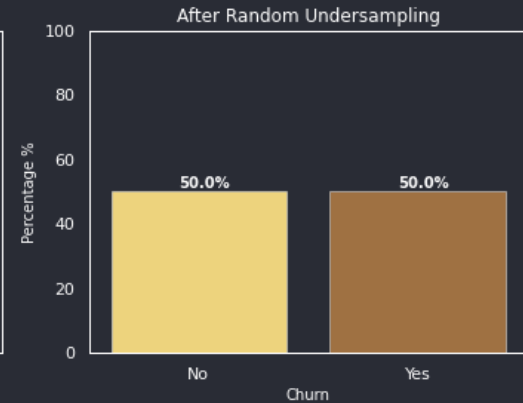
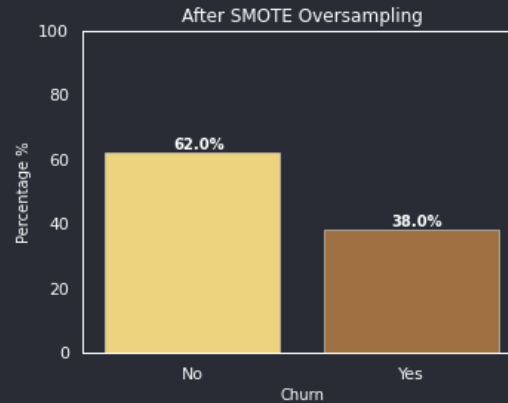
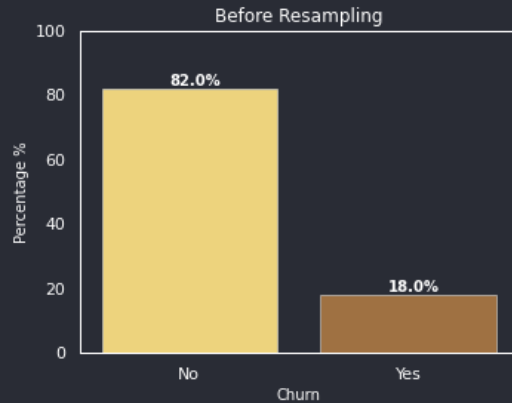
We fitted a **K-means** algorithm with $k = 3$ on the **RFM** table and had the following results:



We concluded to **drop cluster 1**, Thus, **17,949** customers were **excluded** from the study.

Data Preprocessing

To address class **imbalance**, we applied **resampling** techniques as follows:



We also applied a **feature selection** method, and found that **37** is the ideal number of **variables** to keep.

Model selection

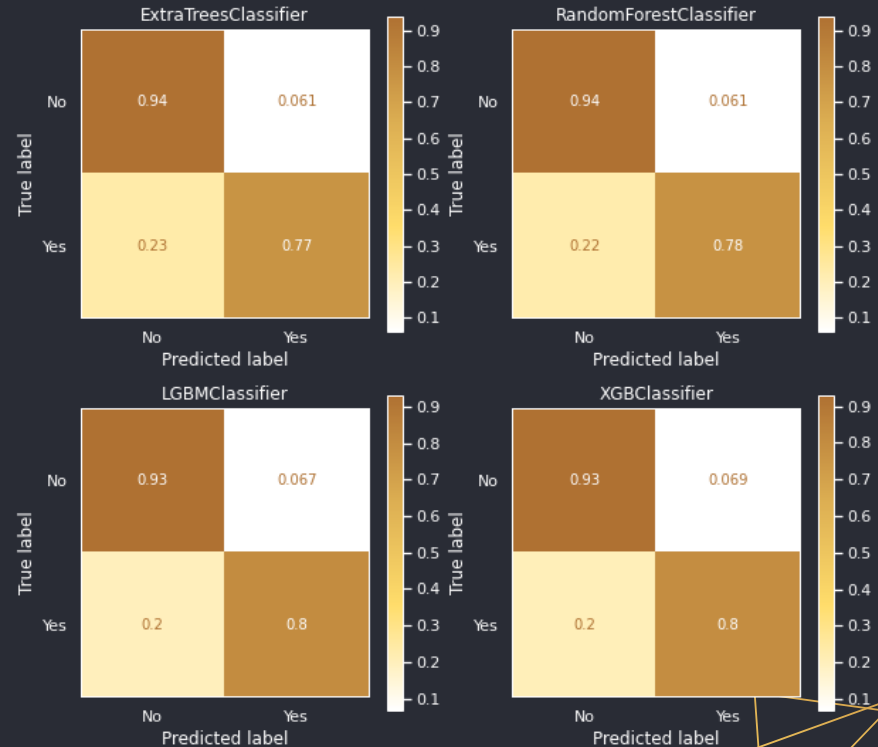
We applied various algorithms on the dataset and received the following results

Name	Accuracy	Recall	Precision	F1_score	Fit Time	Recall_STD
XGBClassifier	0.922	0.917	0.926	0.921	105.999	0.003
LGBMClassifier	0.920	0.914	0.926	0.920	236.458	0.003
RandomForestClassifier	0.912	0.896	0.927	0.911	7.872	0.005
ExtraTreesClassifier	0.912	0.894	0.927	0.910	4.383	0.004
GradientBoostingClassifier	0.907	0.894	0.918	0.906	12.285	0.002
BaggingClassifier	0.905	0.889	0.920	0.904	3.863	0.004
DecisionTreeClassifier	0.874	0.895	0.859	0.877	0.609	0.005
AdaBoostClassifier	0.873	0.843	0.898	0.870	2.966	0.005
LogisticRegressionCV	0.857	0.803	0.902	0.849	18.035	0.003

With ● being the highest value in the columns and ○ being the lowest value

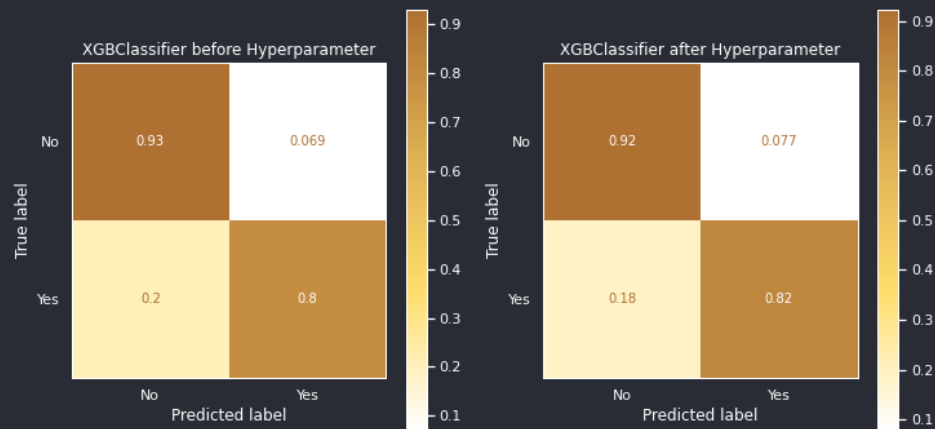
Model selection

- From the previous and the present charts, we choose **XGBClassifier** for validation.



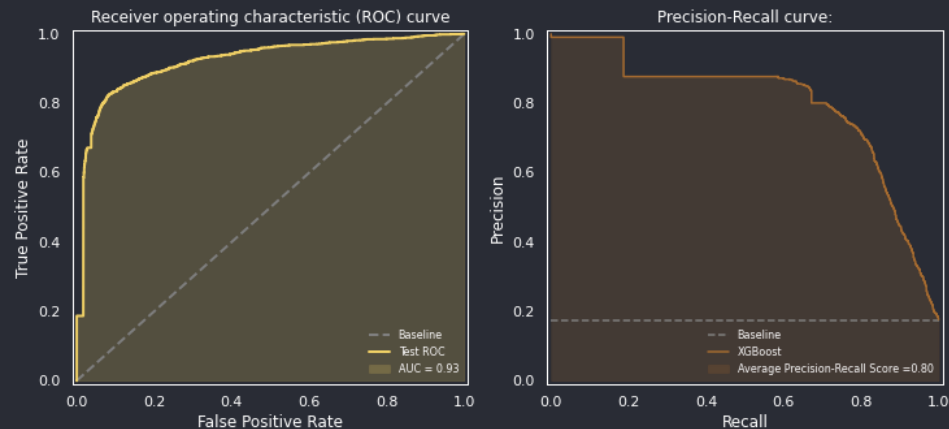
Model validation

Hyperparameter optimization



The new model has **improved** in the **Recall** score of the **positive** class 'Yes'

ROC & Precision – Recall curve



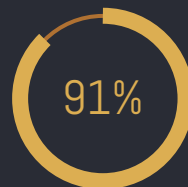
The two curves **satisfy** our **business** requirement

Model evaluation

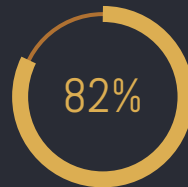
The results of the final model on the test set are as follows:



Accuracy



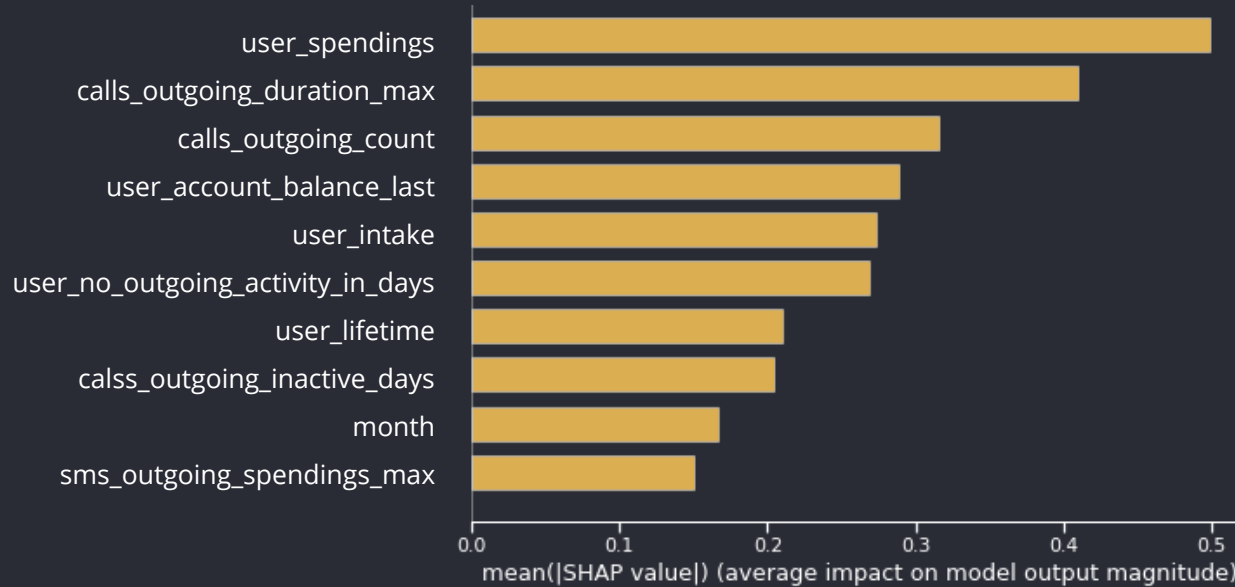
Precision



Recall

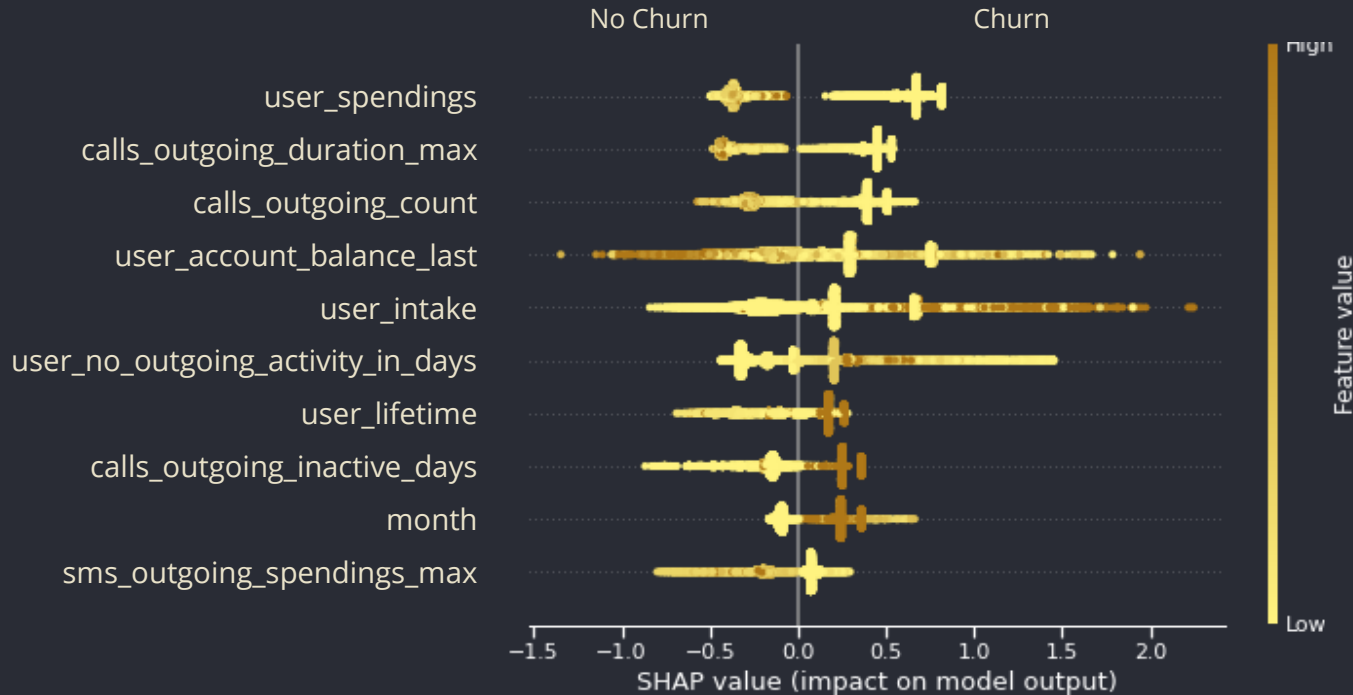
Model Interpretation

Feature importance plot



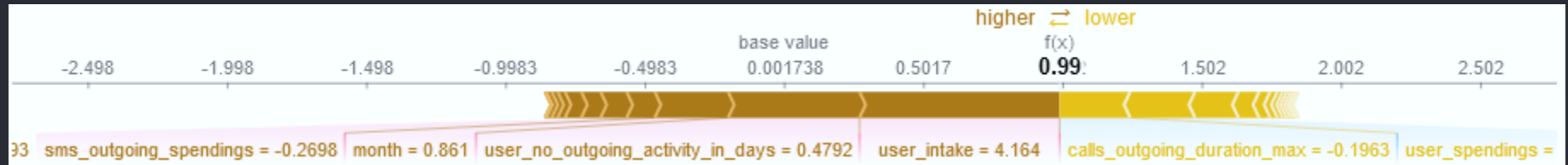
Model Interpretation

Summary plot



Model Interpretation

Individual 1 (Churn)



Individual (Did not Churn)



Conclusion

Based on the **summary** of findings, the following **conclusions** were derived:

01

Ensemble learning algorithms such as Extreme Gradient Boosting (XGBoost) performed the best on the datasets.

02

XGBoost performed better than other models, with an accuracy of 90%.

03

The most significant feature in predicting customer churn is user_spendings.



Recommendations

01

Integrate churn prediction results into the customer relations and marketing departments' IT tools

02

Bring the marketing department closer to CRM and synchronize future actions (retention, promotion, ...).

03

Valorising projects based on data mining to allow the company to align itself with the other competitors



THANKS!

ENSSEA

Author:
YACINE AMMI

Supervisor:
KHALED ROUASKI

CREDITS: This presentation template was created by [Slidesgo](#), including icons by [Flaticon](#), infographics & images by [Freepik](#).