

Customer Churn



“

Customer **churn** is the measure of how many customers stop using a product.

”

churn = change + turn

Why is it so **important** ?

- Acquiring a new customer can **cost up to 5x more** than retaining an existing one.
- Unhappy customers are more likely to **speak negatively** and therefore harm your brand.



10,000

Bank Clients

10 features

Binary target

20% churned

80% retained



Classification Supervised

The **goal** is to build a model that can distinguish customers who are likely to churn.

> 80%

Model prediction score

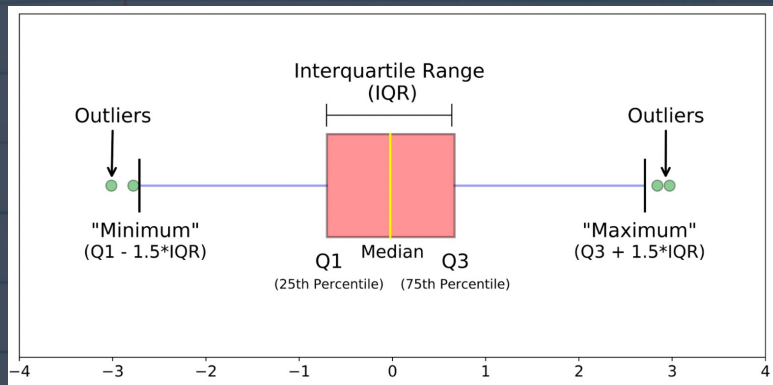


1

Handle Outliers

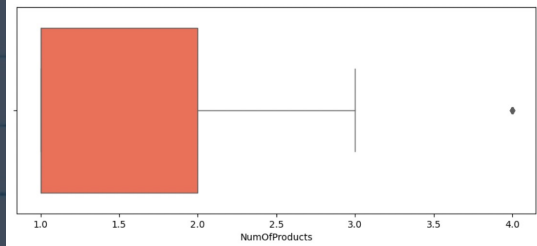
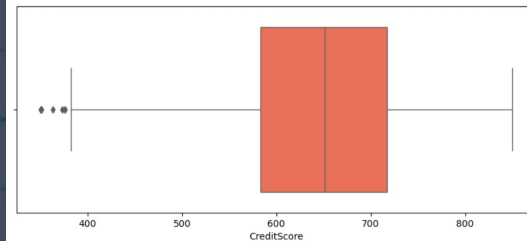
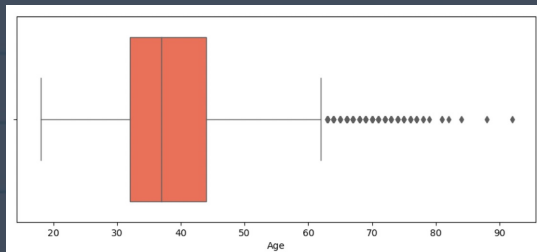
Removed in train so model can generalize better

Interquartile Range (Tukey's rule)



4% removed

5,740 train dataset > 1,000



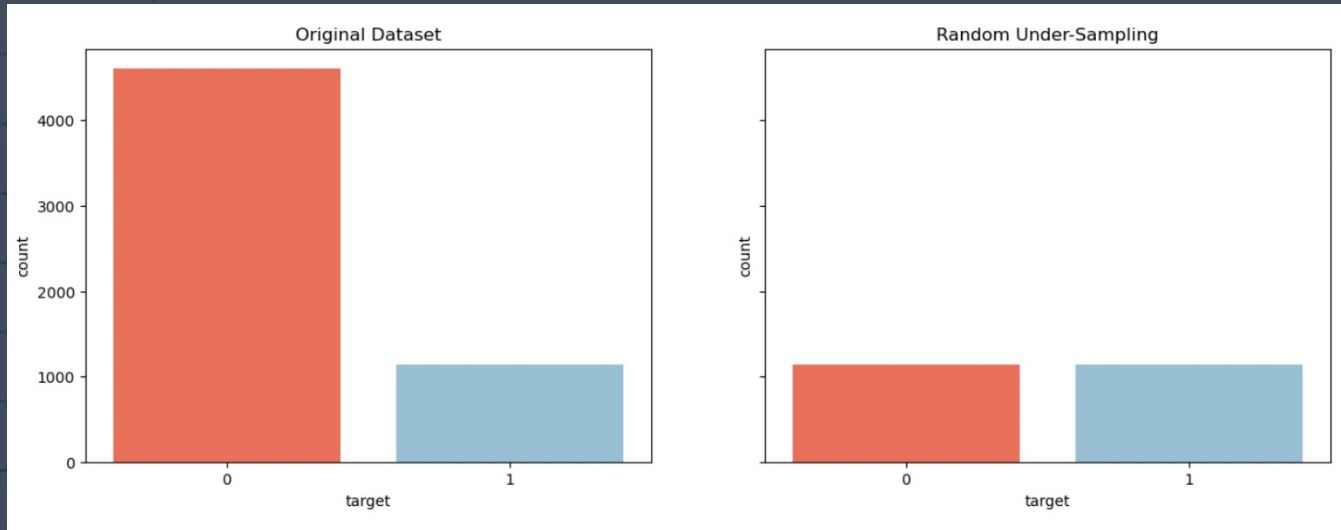


2

Random Under-Sampling

Even up dataset of 80% retained vs 20% churned

Slightly > 1,000

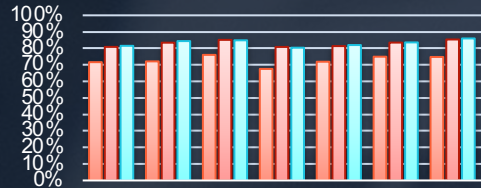




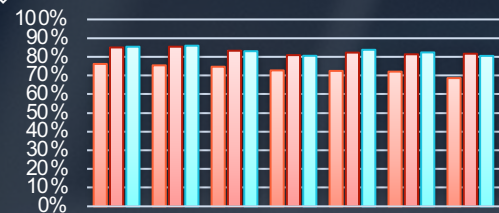
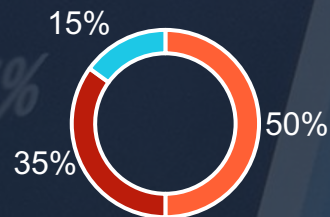
3

Evaluating our models

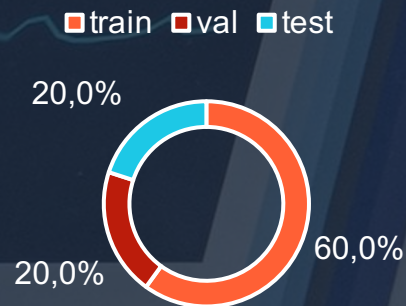
Try different train, validation and test % samples.



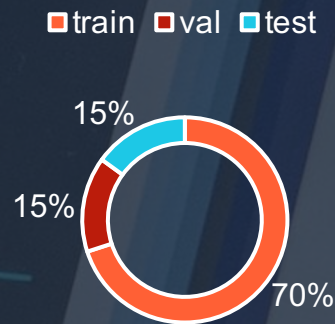
50_35_15	Train	Val	Test
Random Forest	76,05%	85,14%	84,87%
Gradient Boosting	74,95%	83,49%	83,60%
XGBoost	74,79%	85,37%	86,00%
Decission Tree	72,11%	83,43%	84,40%
Naive Bayes	71,90%	81,37%	82,00%
Logistic Regression	71,64%	80,89%	81,53%
kNN	67,54%	80,86%	80,40%



60_20_20	Train	Val	Test
Random Forest	76,18%	85,00%	85,40%
XGBoost	75,44%	85,40%	85,95%
Gradient Boosting	74,69%	83,25%	83,00%
Logistic Regression	72,72%	80,90%	80,55%
Decission Tree	72,46%	82,25%	83,75%
Naive Bayes	72,07%	81,35%	82,35%
kNN	68,61%	81,65%	80,50%

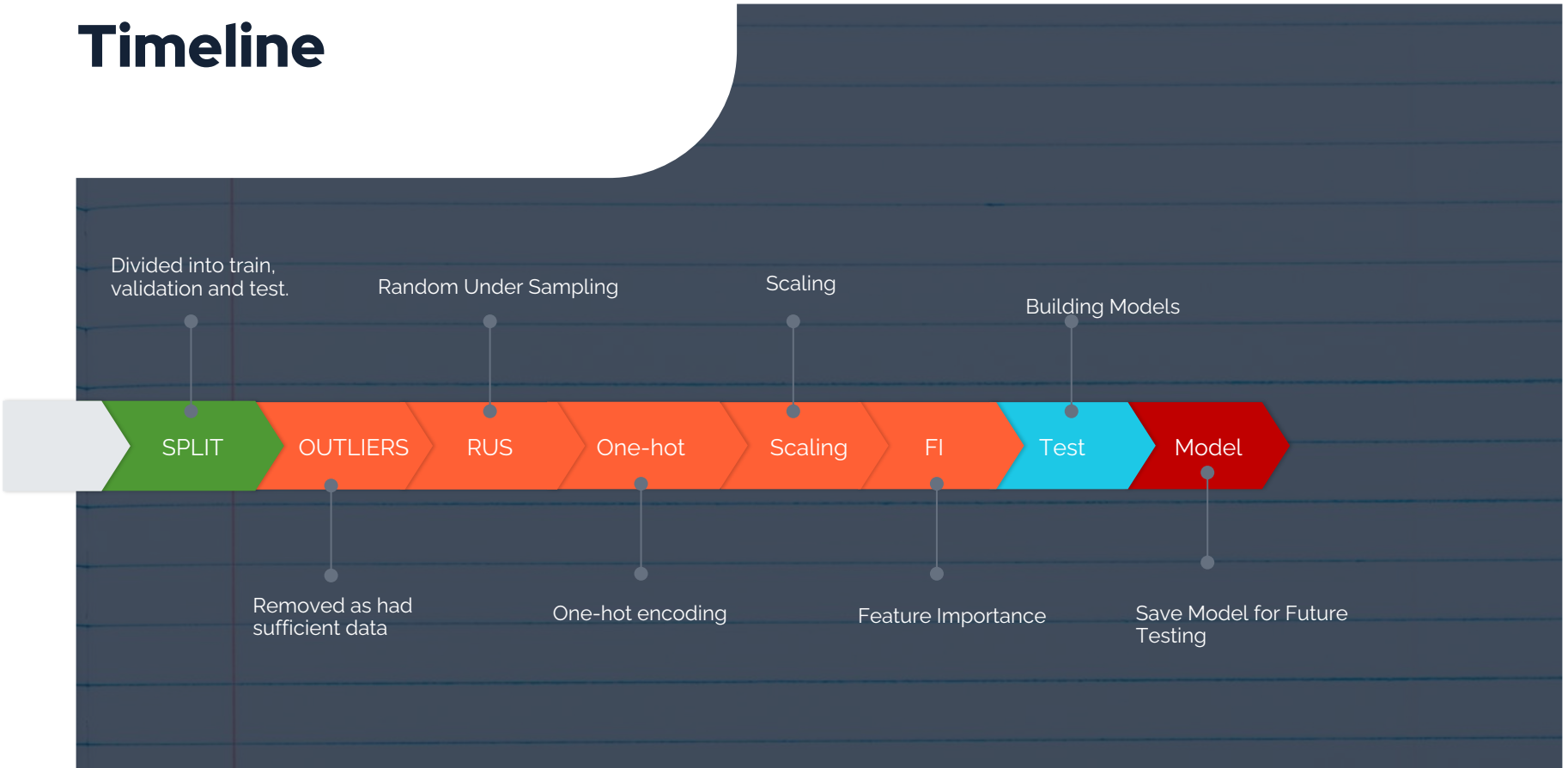


70_15_15	Train	Val	Test
Random Forest	77,06%	85,40%	85,67%
XGBoost	76,83%	85,60%	84,93%
Gradient Boosting	75,38%	83,40%	82,93%
Logistic Regression	72,41%	79,47%	80,40%
Naive Bayes	71,65%	81,53%	80,73%
Decission Tree	71,27%	83,80%	82,73%
kNN	69,83%	79,40%	79,13%



train val test

Timeline



Thank you!

I am Ignacio García-Barrero

I am a Data Science student

You can find me at

github.com/IgnacioGB1990

