INTRODUCTION

In a telco company, there are two costs known as Acquisition Cost and Retention Cost. Acquisition Cost is the expense for a company to acquire new customers. Meanwhile, Retention Cost is the spending for the company to retain existing customers.

Due to human limitations, we often make errors in predicting which customers will churn and which will remain. Consequently, fund allocation can be inaccurate, resulting in excessive expenditures.

Furthermore, according to various sources, the acquisition cost is seven times higher than the retention cost. If we make a mistake in predicting those customers we end up spending more than necessary.

Proposed OBJECTIVES

In this project, our goal is to create a smart computer program using Machine Learning that can guess which customers might decide to stop using our services and which ones will continue using them. The idea is to make this guessing game as accurate as possible. Why? So we can be really smart about where we spend our money on trying to keep customers and where we don't need to spend as much.

Objective 1

Identify the factors influencing customer churn.

Objective 2

Create a machine learning model capable of predicting churn.

Objective 3

Minimize costs.

Assumption

$10

Retain Cost

Retention cost, also known as Customer Retention Cost, refers to the expenses a company incurs to retain its existing customers and prevent them from leaving or churning.

$70

Acquisition Cost

Acquisition cost, often referred to as Customer Acquisition Cost (CAC), is the expense incurred by a company to acquire new customers. It represents the cost associated with convincing a potential customer to make their first purchase or start using a company's products or services.

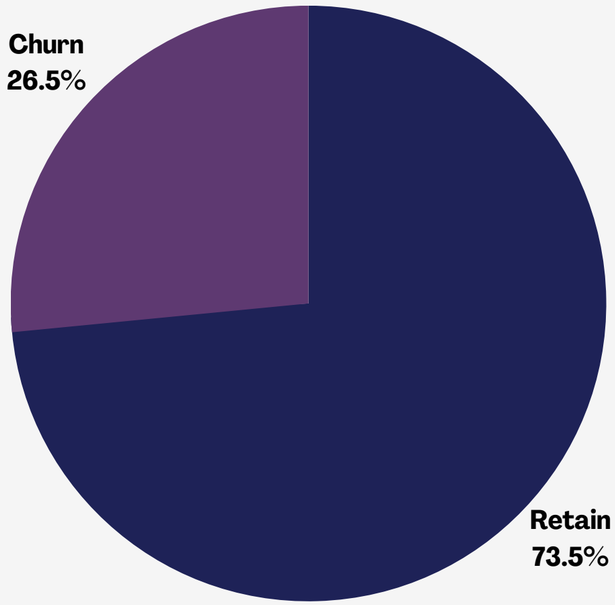
THE DATA

The dataset we are using is a fictional telco company that provided home phone and Internet services to 7043 customers in California in Q3.

7043 observations with 33 variables

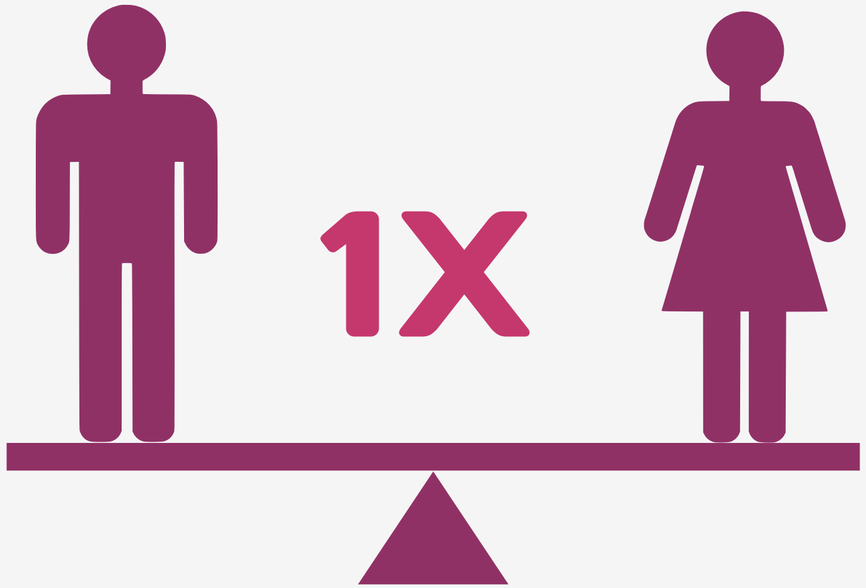
Telco customer churn (11.1.3+)

community.ibm.com / Jul 11, 2019



Churn PROPORTION

From the pie chart above, we can see that **26.54% of the customers in this dataset are labelled as churn customers**. This indicates that there is an imbalance between the number of customers leaving and those staying. While some might think that this dataset needs to be resampled or, in other words, the labels should be "balanced," I am hesitant to take that step at this point. Instead, I will conduct further exploration to gain more insights.



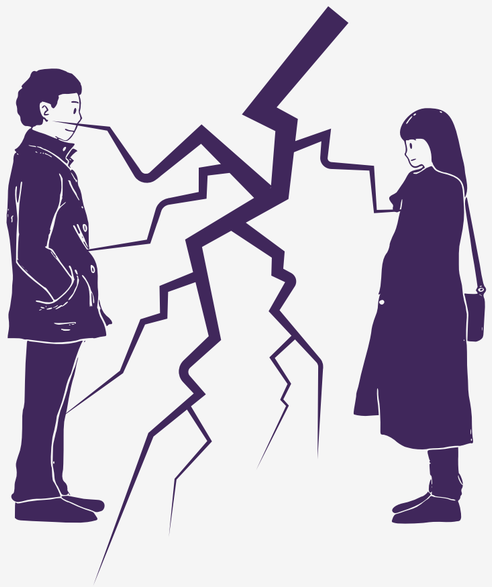
Churn
BY GENDER

It can be observed that the probability of churn based on gender **does not differ significantly** between men and women.

Churn
BY SENIORS

↑1.7X

The elderly have **nearly twice the chance of churning** compared to the younger generation.



1.7X

Churn
BY PARTNER

Customers without a partner also have a tendency to **churn almost twice** as much as those who have a partner.

5++X

Churn
BY DEPENDENTS

Customers who have children (or other dependents) have a **5 times greater chance of churning** compared to customers who do not have them.

↑6X

Churn

BY INTERNET SERVICE SUBSCRIPTION

Customers who use a **fiber optic connection** for their internet service subscription have **nearly 6 times greater chances of churning** compared to those who do not subscribe to internet services.

5++X

Churn
BY DEVICE PROTECTION

Customers who **do not subscribe to device protection** have a tendency to **churn more than 5 times** that of those who use device protection.

15X

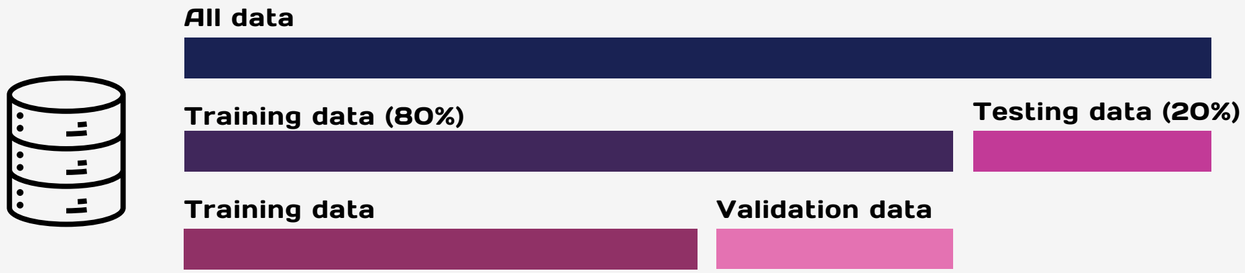
Churn
BY CONTRACT

Customers with **month-to-month contracts** are **15 times more likely to churn** than customers with 2-year or annual contracts.

3X

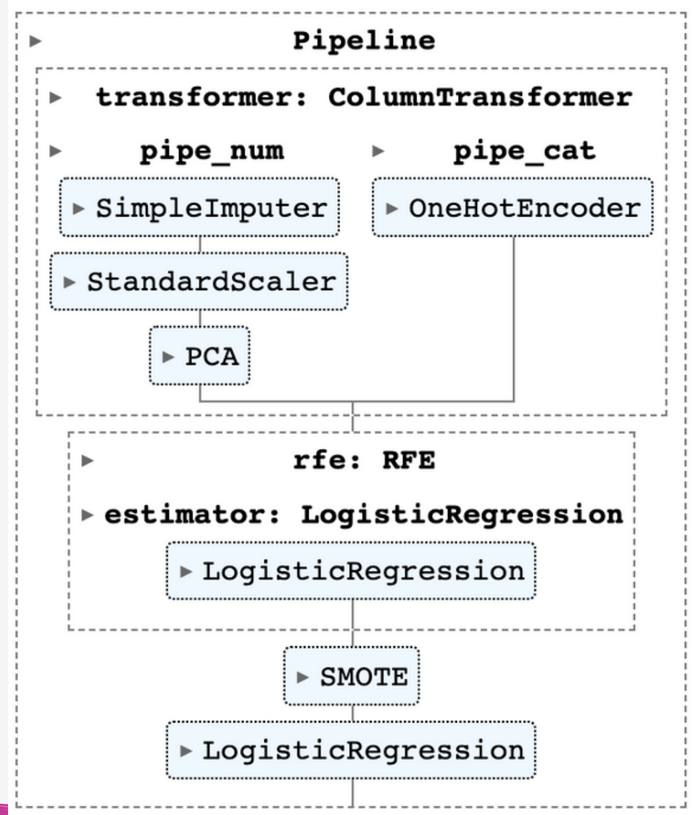
Churn
BY PAYMENT METHOD

Customers who use **electronic checks as their payment method** are **approximately 3 times more likely to churn** compared to other payment methods.



Modelling STRATEGIES

The strategy is, first, to **divide the data into two** parts: **Training Data (80%)** and **Testing Data (20%)**. From the Training Data, we will further split it into two parts: Training Data and Validation Data. The Validation Data will be used to assess the model's predictive abilities.

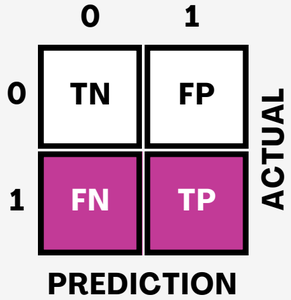


Modelling STRATEGIES

The second step involves **creating a Pipeline** that allows us to process the data up to the model fitting stage. Why use a Pipeline? This is to **avoid data leakage.**

Data leakage occurs when information from the **testing dataset "leaks" into the training process**. In other words, the model unintentionally gains access to information it shouldn't have during training, which **can lead to unrealistically optimistic performance estimates**.

Recall



$$\frac{TP}{TP+FN}$$

Modelling STRATEGIES

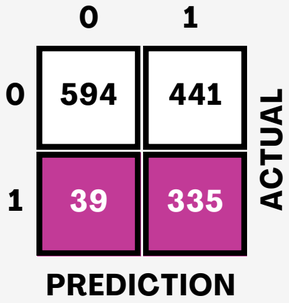
We will attempt to **optimize the Recall value**. Referring to our main objective, **which is to minimize costs**, and considering that the largest cost comes from acquisition cost, **we aim to minimize the possibility of customers who are likely to churn but do not receive any form of compensation to stay** (retain cost).

Modelling
RESULT

Recall

89%

Precision*
43%



We have successfully maximized the **Recall score to 89%**.
However, what does this mean for the business?

***We will discuss it later**

1000 Customers that actually churn

With Machine Learning (89% Recall)

890 Number of customers we correctly predict as churn

110 Number of customers we misclassified as retain customers

Breaking Down the Cost

$8,900 Retain cost (@ $10)

$7,700 Acquisition cost (@ $70)

+

$16,600

Total cost

Result
INTERPRETATION

If we have **1000 customers with the potential to churn**, then we will accurately predict that **890 customers will churn**. Thus, we can allocate a retaining **cost of $8,900** (assuming the initial retention cost is $10 per customer). However, there are **110 customers we did not predict accurately**, so we need to spend an acquisition cost to replace those lost customers, **which amounts to $7,700** (assuming the initial acquisition cost is $70 per customer). The **total cost** we incur is **approximately $16,600**.

What if we didn't use machine learning?

1000 Customers that actually churn

Without Machine Learning (50% Chances)

500 Number of customers we correctly predict as churn

500 Number of customers we misclassified as retain customers

Breaking Down the Cost

$5,000 Retain cost (@ $10)

$35,000 Acquisition cost (@ $70)

+

$40,000

Total cost

Result INTERPRETATION

Without using machine learning, the most **optimistic guess would have an accuracy of 50%**. This means we would accurately guess **500 customers who will churn**. Thus, we can allocate a **retaining cost of $5,000** (assuming the initial retention cost is $10 per customer). However, there are **500 customers we did not guess correctly**, so we need to spend an acquisition cost to replace those lost customers, **which amounts to $35,000** (assuming the initial acquisition cost is $70 per customer). **The total cost** we incur is **approximately $40,000**.

$16,600 - $40,000

-$23,400

(-59%)

Result
INTERPRETATION

Therefore, in this case, using Machine Learning can help us **save costs by up to 59%**.

2070*

Customers that predicted churn

*From 890 x 100/40

With Machine Learning (43% Precision)

890 Number of customers we classified as churn and truly churn

1180 Number of customers we misclassified as churn customers (They actually retain)

Breaking Down the Cost

$20,700 Retain cost estimated (@ $10)

$11,800 Retain cost off-target (@ $10)

$8,900

Retain cost on-target

Result

INTERPRETATION (PRECISION)

Out of the **2070 customers we predicted as churn**, we only **accurately predicted 890** who actually churned. The rest are customers who were retained but predicted as churn. So, we still need to allocate retention cost for those customers. In other words, out of the **$20,700** we spent on retention cost, **$8,900 hit the mark**, and **the rest did not**.

ACTUAL \ PREDICTION	0	1
0	594	441
1	39	335

Precision

43%

$16,600 (+ $11,800) - $40,000

-$11,600

(-29%)

Result
INTERPRETATION

Therefore, if we take into account the Precision score, in this case, using Machine Learning can help us **save costs by up to 29%**.