TELCO CHURN PREDICTION

BINARY CLASSIFICATION

MACHINE LEARNING PORTFOLIO

DEREK RODRIGO SÁNCHEZ SEGUAME

1. Task Definition
2. EDA
3. Model and Results

01

Task Definition

PROBLEM STATEMENT | CUSTOMER CHURN IN TELCO

TeleConnecta Solutions is a new telecommunications provider in Mexico that has successfully achieved its initial market objectives. Their main focus is individual customers and the've succesfuly executed an aggressive market expansion strategy.

However, the company now faces a higher-than-expected **churn rate**, directly affecting profitability and competitiveness. To tackle this challenge, TeleConnecta plans to implement a **Machine Learning model** as part of a **data-driven strategy**. This approach will enable the organization to identify at-risk customers, optimize retention efforts, and personalize offers—all **powered by AI** and built on the pillars of **reproducibility, scalability, and data quality**.

The available data for this project consists of the Telco Customer Churn dataset. By applying advanced analytical techniques, TeleConnecta aims to **reduce churn**, ensure **regulatory compliance**, and drive **sustainable growth** in Mexico's competitive telecommunications market.

COMMUNICATIONS AND CUSTOMER CHURN | RISK MANAGEMENT POWERED BY MACHINE LEARNING

Our client

- **TeleConnecta** is a mexican Fintech
- Their target market are **SMEs**
- They've had an expansive strategy to **achieve market share** goals

Pain point

- they are experiencing **excessive defaults** due to:
  - Its broad market reach
  - **Unsound retention strategy**
- This generates problems in:
  - Profitability
  - Competitveness

Solution

- Using the provided dataset, a **Machine Learning Binary Classification Model** will be generated
- This model will evaluate each client and **estimate the churn possibility**

Note:

02

Exploratory Data Analysis

OVERVIEW | A DATASET PRIMARILY COMPOSED BY CATEGORICAL FEATURES

of Features | By data type

	#Features
Numerical	3
Categorical non-ordinal	19
Total	**21**

# of Categorical Features | By # of unique observations

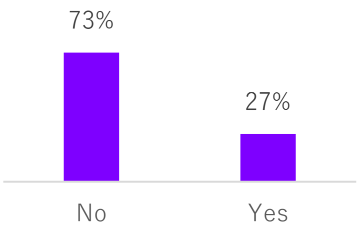
	Two Unique Obs	Three Unique Obs	Four+ Unique Obs
#Features	7	9	3

Categorical Features Distributions

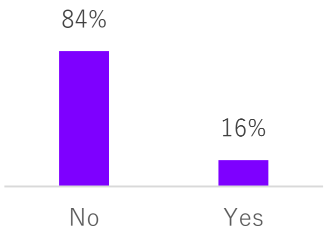
Target

Features

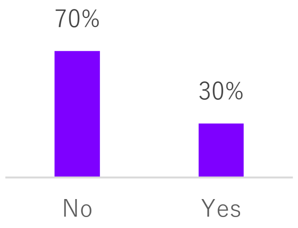
Churn



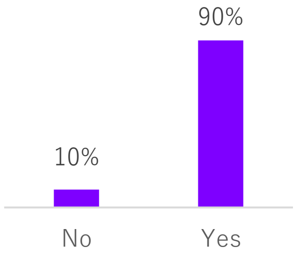
Senior



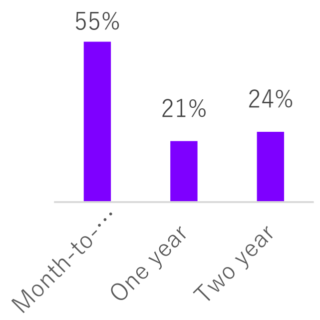
Dependents



Phone



Contract



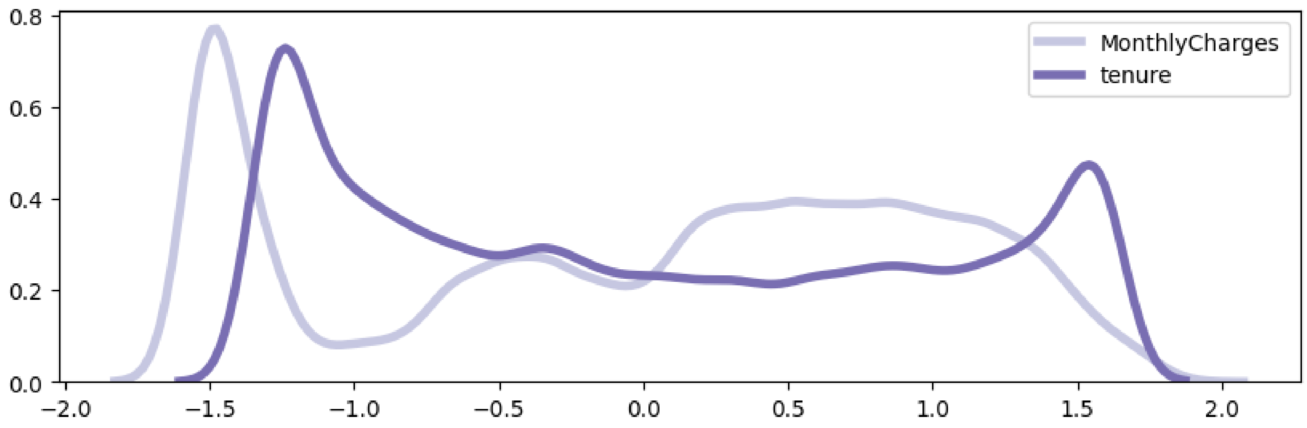
Note:

NUMERICAL VARIABLES | HEATMAP, DISTRIBUTIONS AND BOXPLOTS

Correlation Matrix Heatmap

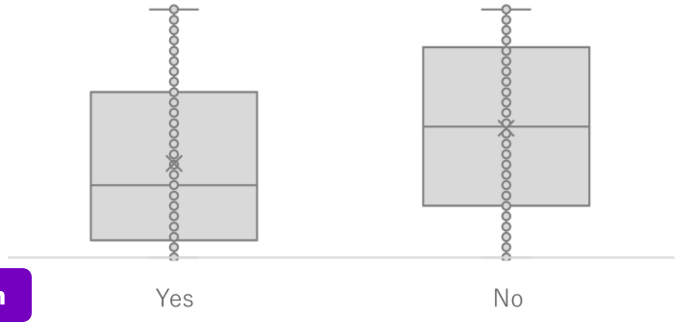
	Churn	Monthly	Total Charge	Tenure
Churn	1	0.13	-0.11	-0.2
Monthly	0.13	1	0.65	0.24
Total Charge	-0.11	0.65	1	0.82
Tenure	-0.2	0.24	0.82	1

Distribution Graph

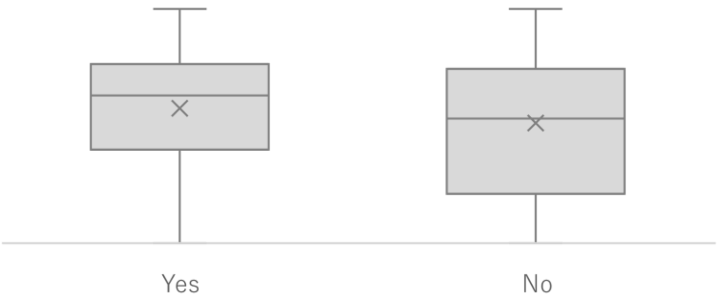


Box plot | by feature

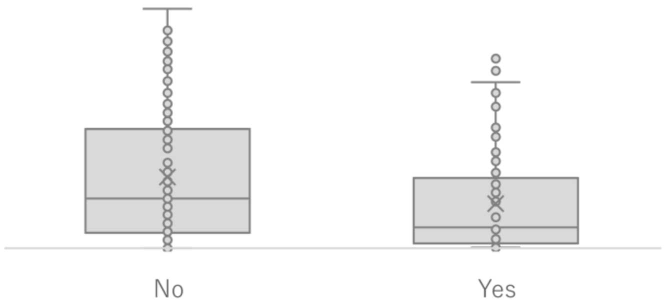
Tenure



Monthly



Total Charge



Churn

03

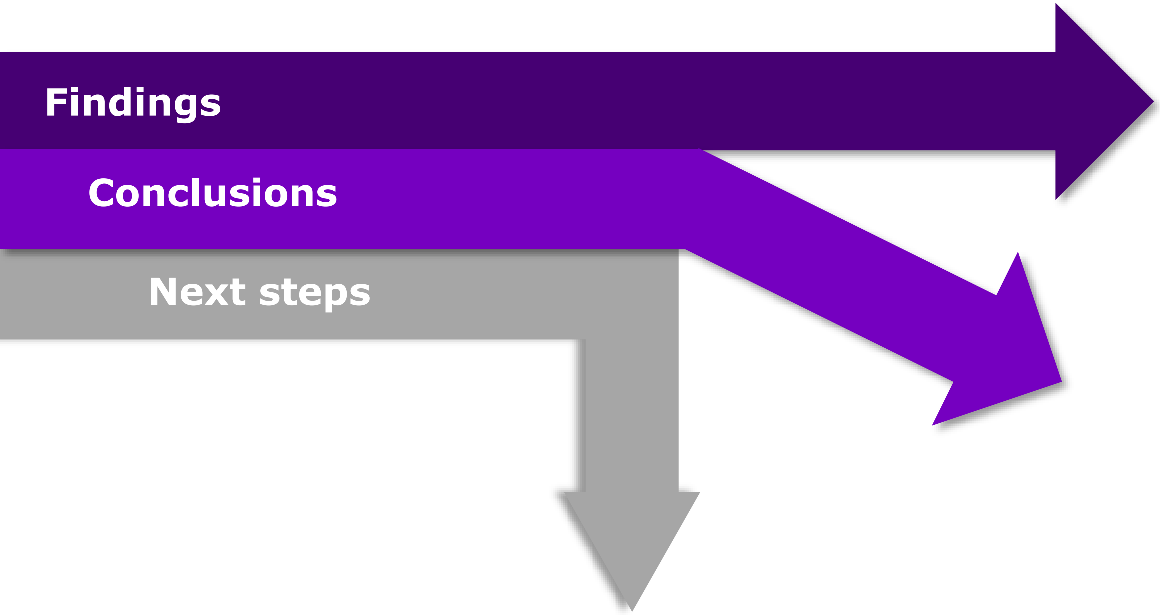
Model Selection and Results

MODEL SELECTION | GRADIENT BOOSTING A LOGISTIC REGRESSION PERFORMED ROUGHLY THE SAME BUT BETTER THAN K-NN

Model	Tuning	Parameters	Metrics
Gradient Boosting	Randomized Search	learning rate = 0.01 n_estimators = 100 Max_Depth = 4	**Accuracy: 76%** Precision: 73% AUC ROC: 83%
Random Forest	Randomized Search	n_estimators = 427 max_depth = 17	**Accuracy: 78%** Precision: 67% AUC ROC: 84%
SVM	Grid Search	C = 1.0 kernel = rbf	**Accuracy: 78%** Precision: 66% AUC ROC: 79%
Logistic Regression	Grid Search	L1 Regularization C = 5.2	**Accuracy: 79%** Precision: 66% AUC ROC: 84%

Note:

FROM DATA TO ACTIONABLES | THE POWER OF DATA TO PROVIDE BUSINESS SOLUTIONS



Gradient Boosting has the best precision, but…

- **Simplier models** like Logistc Regression adhere more to regulation and **have easier interpretations**
- Considering the client's concern of regularatory standards, **Logistic Regression might be suitable**

Final Thoughts

- **Newly acquired** customers have the highest **possibility of churn**
- It is important to have a **reliable risk model**
- Despite several tests and treatment, **dataset might be noisy and affect predictive power**

Improving model performance

- Discovery process with the client to ask for a dataset with:
  - **More observations**
  - **More features**