Telco Customer Churn Analyis

Ironhack Mid-Project
David Arques
June 2024

Agenda

Project Overview:
- Goal & Context
- Problem Statement

Data Exploration:
- Demographics
- Customer Segmentation
- Billing

Correlations:
- Bivariate Analysis

Machine Learning

Conclusions

Project Overview

Goal

Analyze the Telco Customer Churn dataset to gain **valuable insights** and identify specific **business actions to reduce customer churn**.

Context

Customer churn = Customer leaving
Retaining existing customers is more important than acquiring new ones
(Product-Led mindset)

Problem Statement

How can we **reduce customer churn using data-driven insights** and action-oriented analysis?

Dataset 💾

Kaggle: Telco Customer Churn
+7K customer and 20 Features

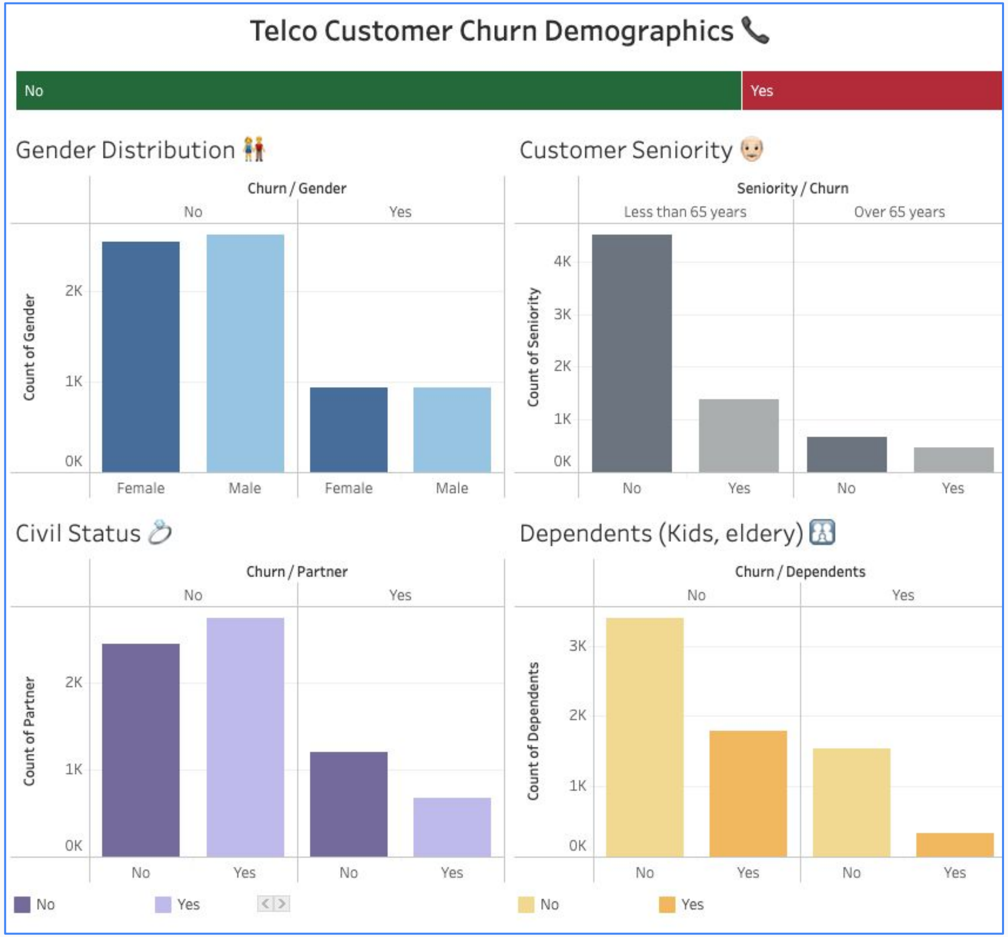
Demographics 🧑‍🤝‍🧑	Service/Tech 📡	Billing 💵	Numerical 📈	Target 🎯
Gender	Phone Service	Contract type	Monthly charges	**Churn**
Seniority	Phone lines	Paperless Billing	Total Charges	Yes = leaving
Partner	Internet Service	Payment Method	Tenure	No = staying
Dependents	Online Security			
	Online Backup			
	Device Protection			
	Tech Support			
	Streaming TV			
	Streaming Movies			

Source: [Kaggle Dataset link]

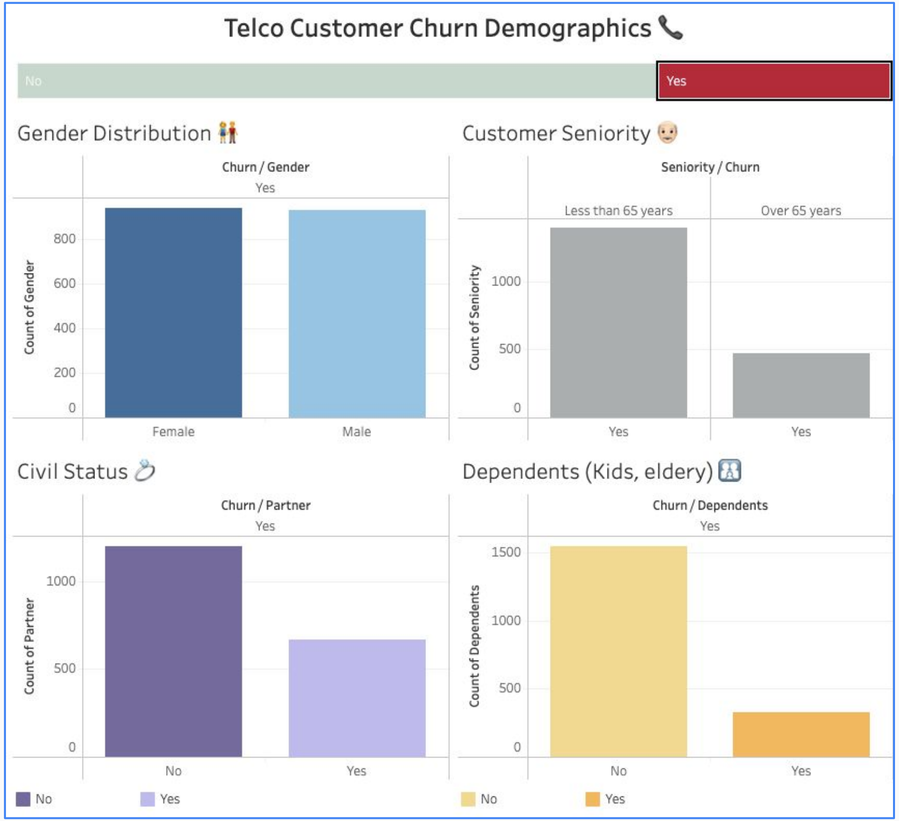
Data Exploration (EDA)

Exploration: Demographics

Interactive customer
[Demographics Dashboard]
by churn



Findings: Demographics 👫



What types of customers are leaving?

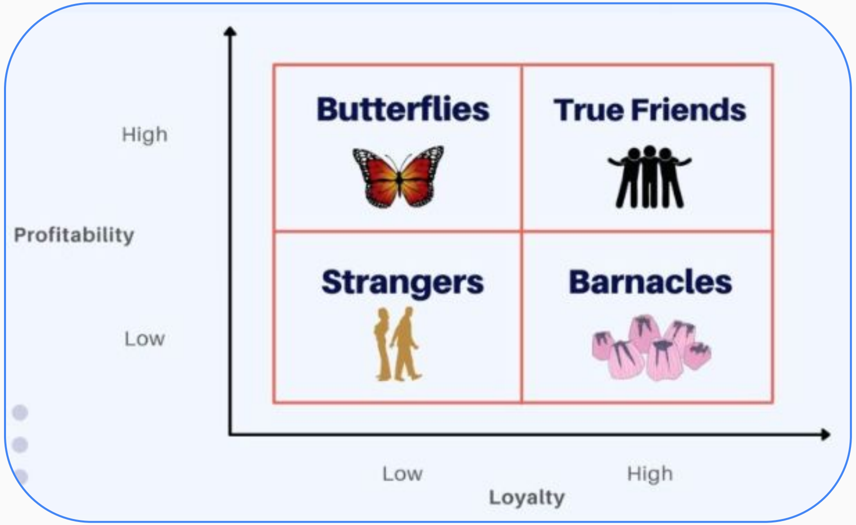
- **Gender** 👫
  The gender distribution is almost **50/50**, so not very relevant.

- **Seniority** 👴
  Most customers leaving are below 65 years old. However, since approximately 80% of our overall data consists of people below 65 (so not relevant).

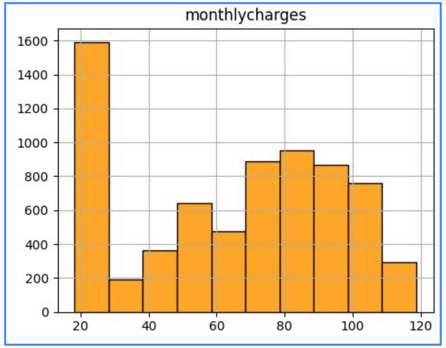
- **Partener/Dependent** 💍 👶
  Most of the customers leaving (around 65%) are single and have no kids or dependents (about 82%).

Action: Focusing on single people with no marriage or kids is a good target audience to improve customer retention.

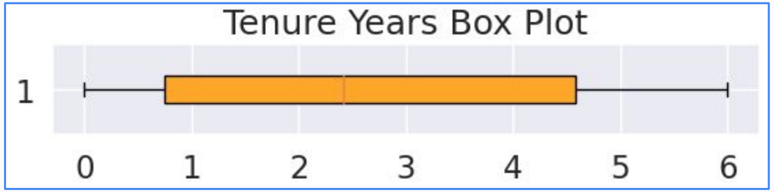
Feature Engineering: Customer Segmentation



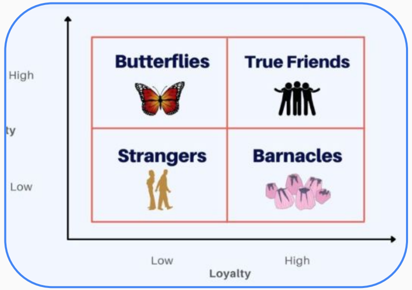
Based on Monthly charges
(mean = 64.8)



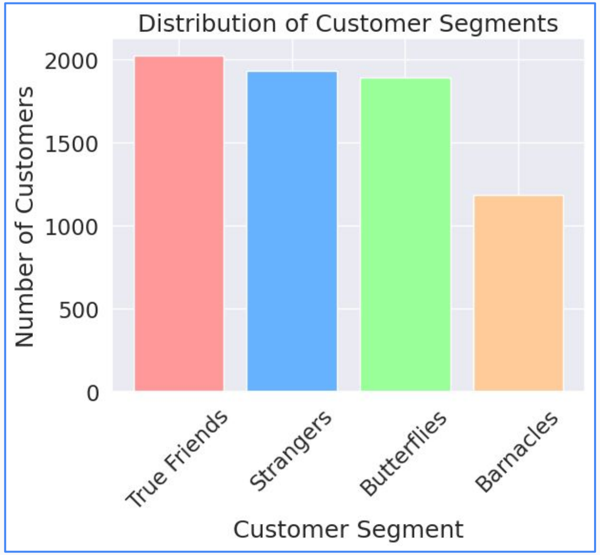
Based on Tenure (mean = 2.7 years)



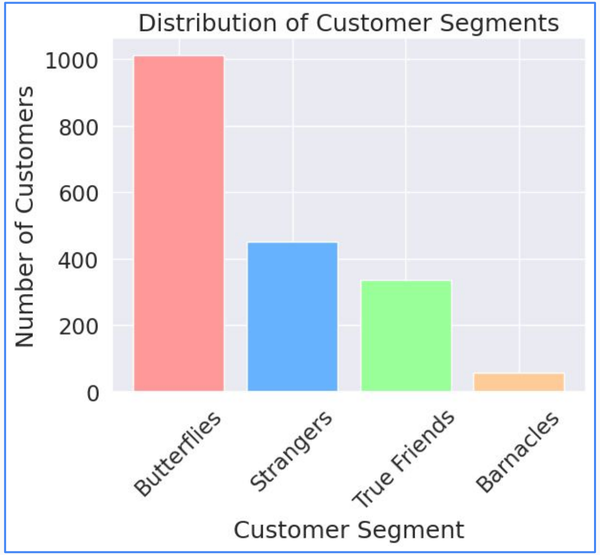
Customer Segmenation Findings



All customer Dataset



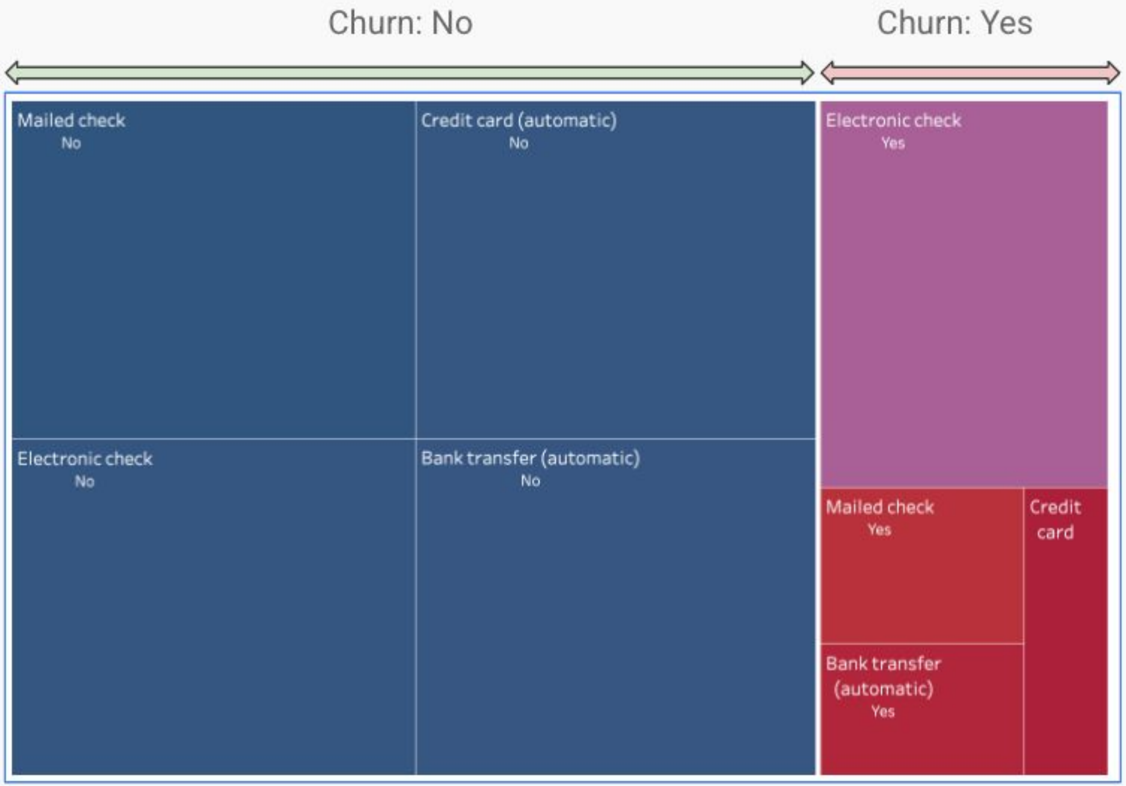
Customers leaving (Churn=yes)



Learning: More than half (~55%) of customers leaving are 'butterflies' 🦋(= highly profitable)
Assessing **customer churn is crucial** to **increasing Telco profit** and improving business.

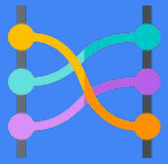
Exploration & Findings: Billing

Insighst: **Review** the **electronic check** payment **method**, as most customers leaving use that one.



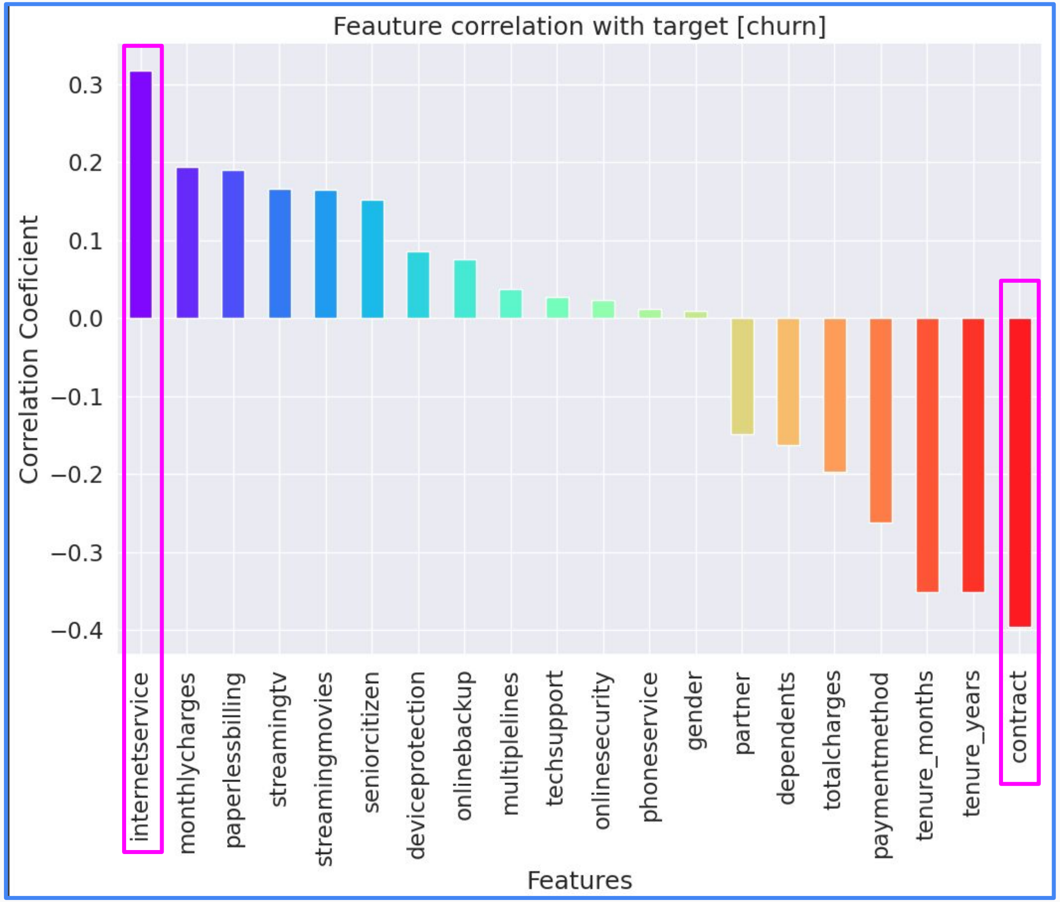
Correlations

Feature correlation assessment with target [churn]



Top correlations to churn:

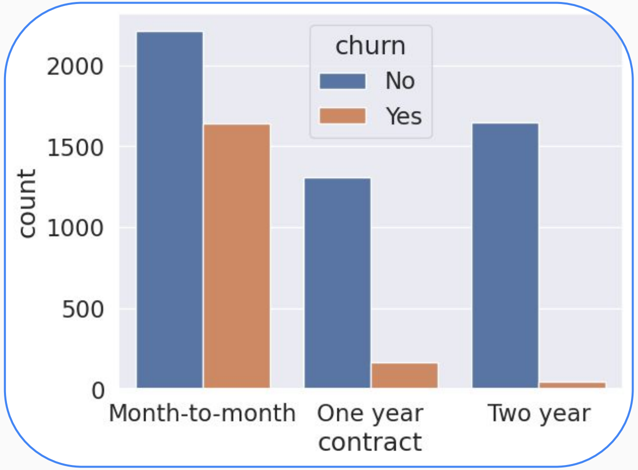
Contract type (~0.40)
Internet service (~0.32)



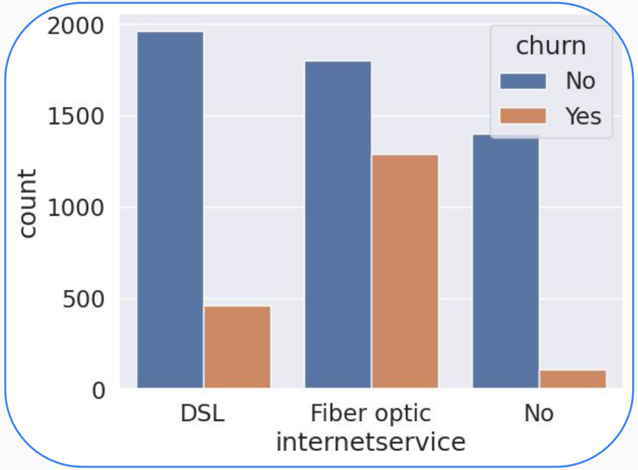
Bivariate Analysis
with High Correlation Features to Target (Churn)

Feature Distribution Visualization

Contract Type

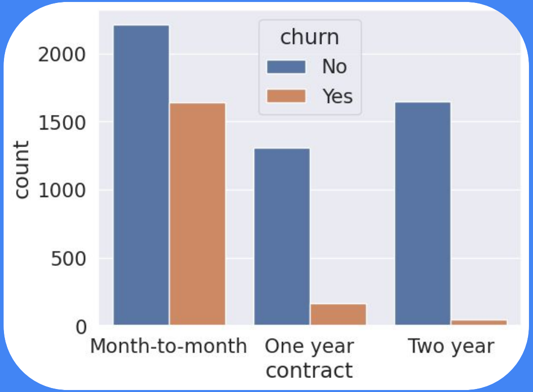


Internet Service



Contract type

Contract type distribution with (~0.40) correlation towards target churn



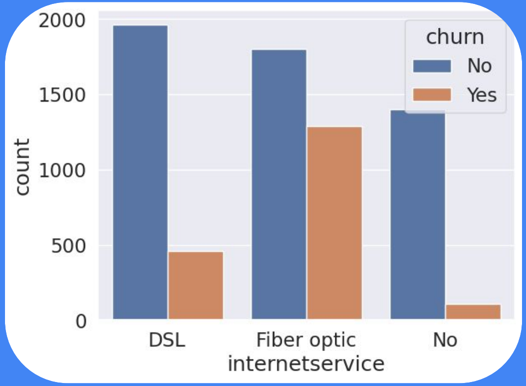
Bivariate Analysis results			
Chi-Square		Cramér's V	
Statistic	1179.80	Association	0.40
p-value	6.44e-257		

Conclusions

- **Chi-square** results show **strong statistical evidence** that contract type is related to customer churn (**p-value < 0.05**).
- **Cramer's V of 0.41** indicates a **moderate relationship**.
- **Contract** type is **not the only driver** of **churn** but significantly influences customer behavior.

Internet Service

Internet Service distribution with (~0.32) correlation towards target churn



Bivariate Analysis results			
Chi-Square		Cramér's V	
Statistic	732.06	Association	0.32
p-value	1.09e−159		

Conclusions

- **Chi-square** results show **moderate statistical evidence** that customer's internet service type is related to churn (**p-value < 0.05**).
- **Cramer's V of 0.32** indicates a **moderate to low relationship**.
- **Customer's internet service** type is **not as strong** as contract type **but is still relevant** when assessing churn.

Machine Learning Testing

ML Data Overview

ML type 🤖

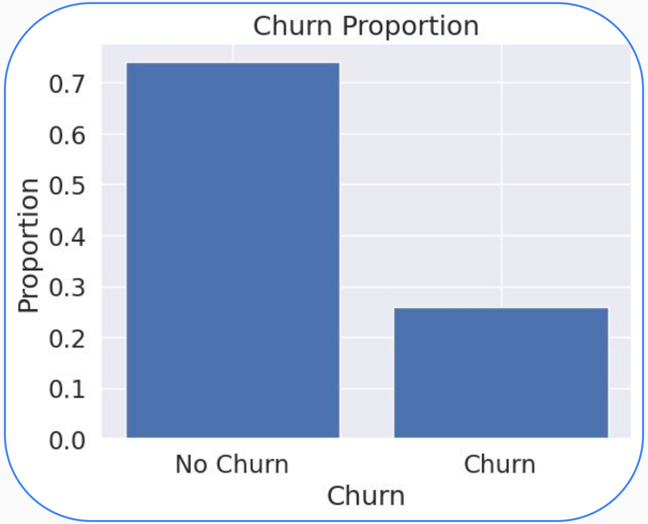
- Classification
- 70/30 Train-test split

Traget 🎯

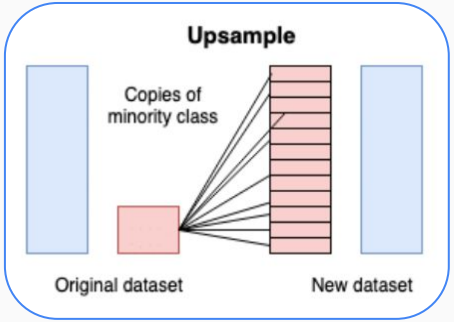
- Churn (No=0, Yes=1)

Size 📏

- 7k Customers (Rows)
- 20 Features (Columns)



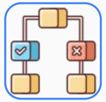
Uneven target data. We'll **Apply Oversampling**



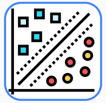
Classification ML models tested

No Data modification

Logistic Regression



Decision Tree Classifier



Support Vector Machine

Upsampling (SMOTE)

Logistic Regression

Random Forest Classifier

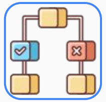
ML Results (No Upsampling)

Test Data accuracy table:	
Logistic Regression	Test: 0.79
	Train: 0.80
Decision Tree	Test: 0.73
	Train: 0.99
SVC	Test: 0.57
	Train: 0.58

Logistic Regression

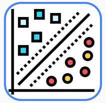
	precision	recall	f1-score	support
0	0.85	0.89	0.87	1556
1	0.63	0.54	0.58	551

Decision Tree Classifier



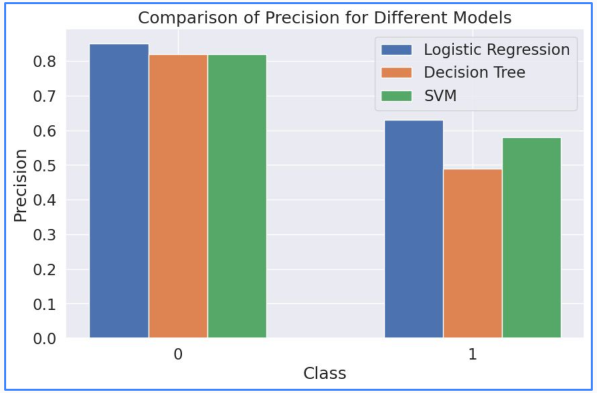
	precision	recall	f1-score	support
0	0.82	0.82	0.82	1556
1	0.48	0.48	0.48	551

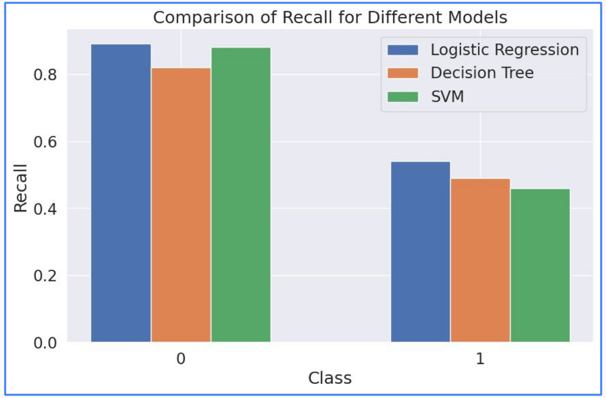
Support Vector Machine (SVC)

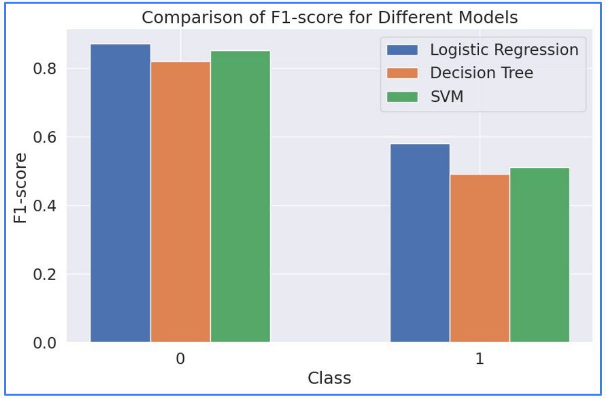


	precision	recall	f1-score	support
0	0.89	0.49	0.63	1556
1	0.36	0.83	0.51	551

ML Results (No Upsampling)





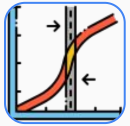


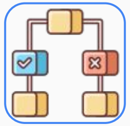
Conclusions

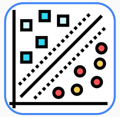
- **Logistic regression** achieved the **best results** across all metrics (Precision, Recall, and F1-score) and Test/Train Accuracy.

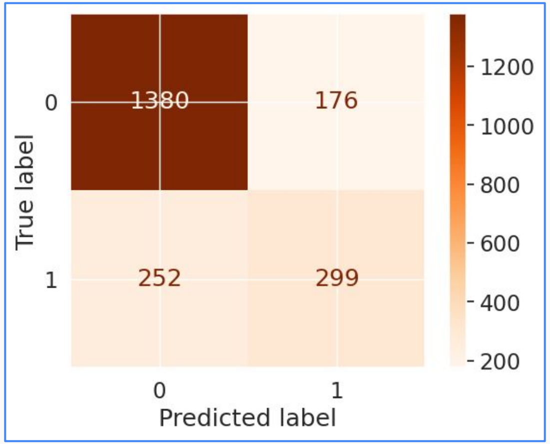
- We can confirm it is the **best ML model** out of the three tested.

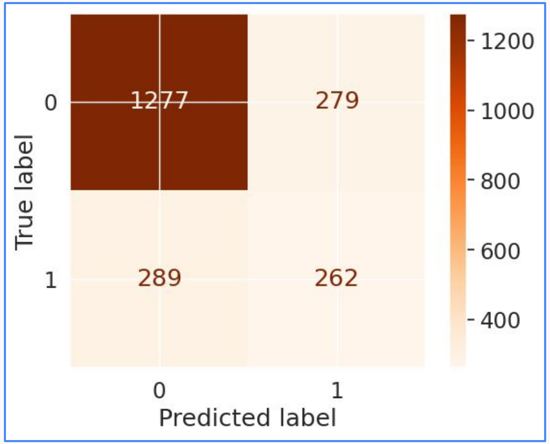
Confusion Metrics (No Upsampling)

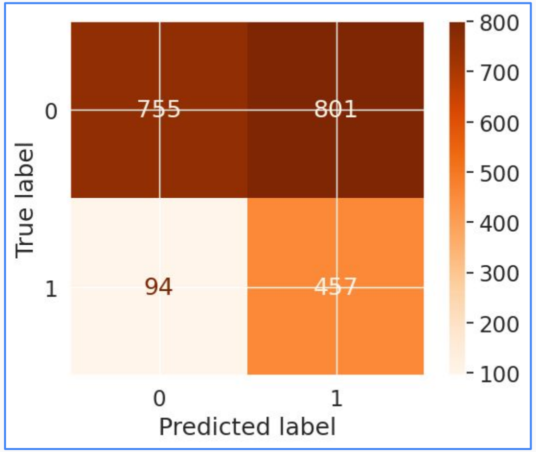












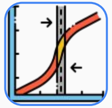
Better model ⭐

ML Results

Upsampling (SMOTE)

Test Data accuracy table:	
Logistic Regression	Test: 0.79
	Train: 0.80
Random Forest	Test: 0.85
	Train: 0.99

Logistic Regression



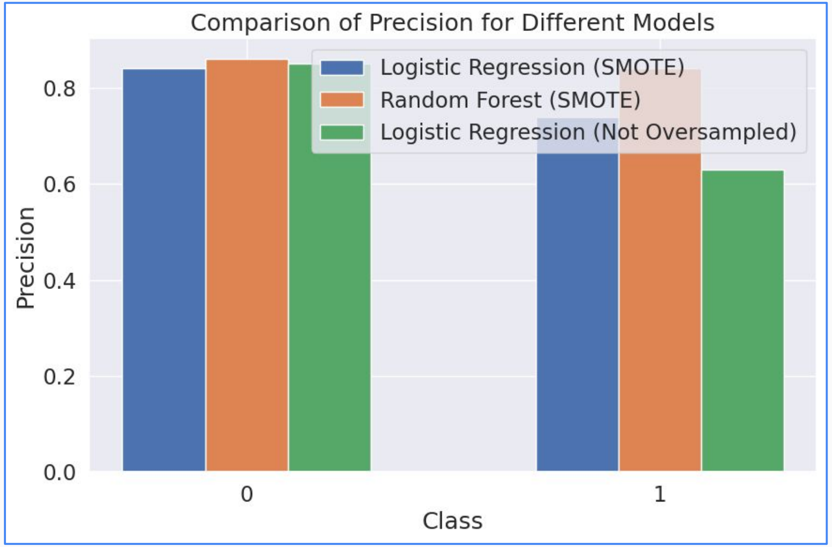
SMOTE

	precision	recall	f1-score	support
0	0.84	0.73	0.78	1586
1	0.75	0.85	0.80	1513

Random Forest Classifier

	precision	recall	f1-score	support
0	0.86	0.84	0.85	1586
1	0.84	0.86	0.85	1513

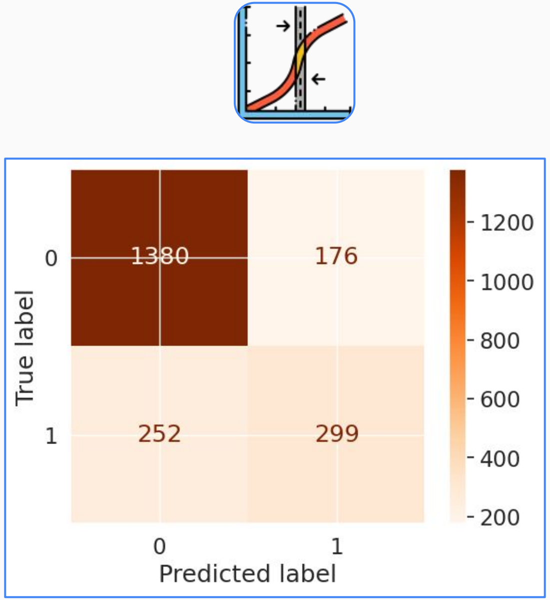
ML Results (No Upsampling)

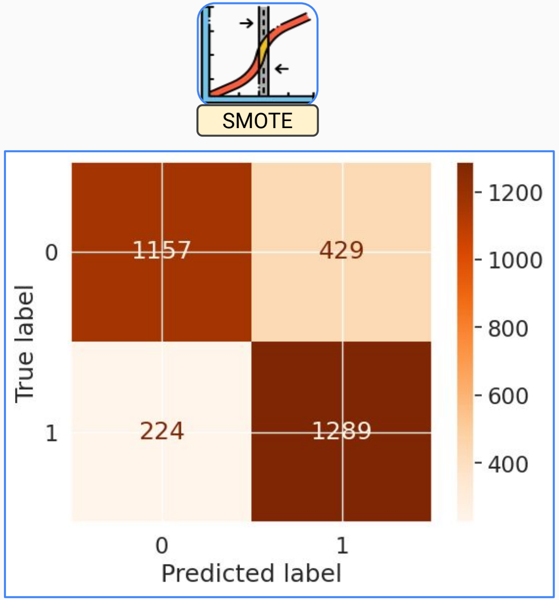


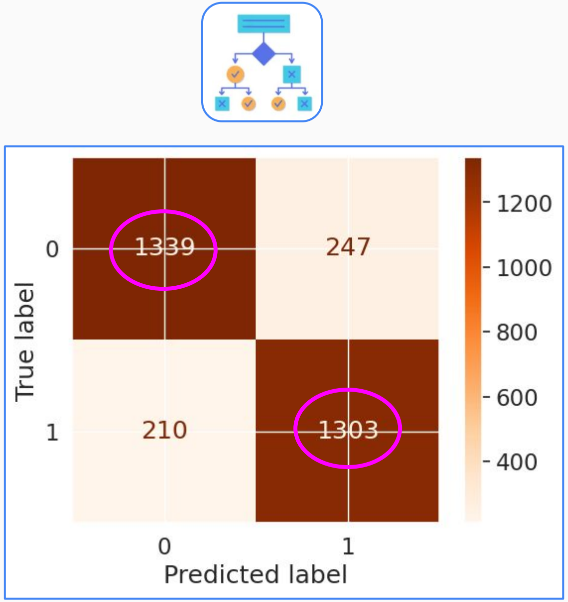
Conclusions

- **After upsampling**, the **best** tested **model** is the **Random Forest** classifier.
- **Precision** metrics **increased** significantly:
  - 0.86 for "Churn No"
  - 0.84 for "Churn Yes"
- Therefore, it is a **better model**.

Confusion Metrics (No Upsampling)

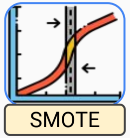


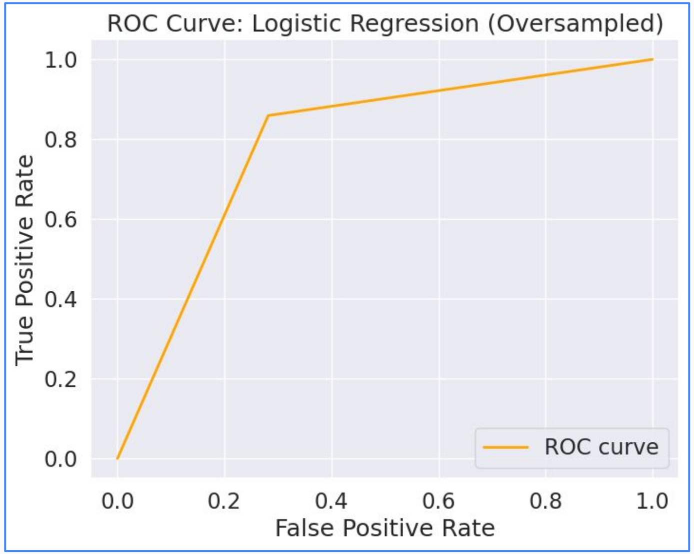


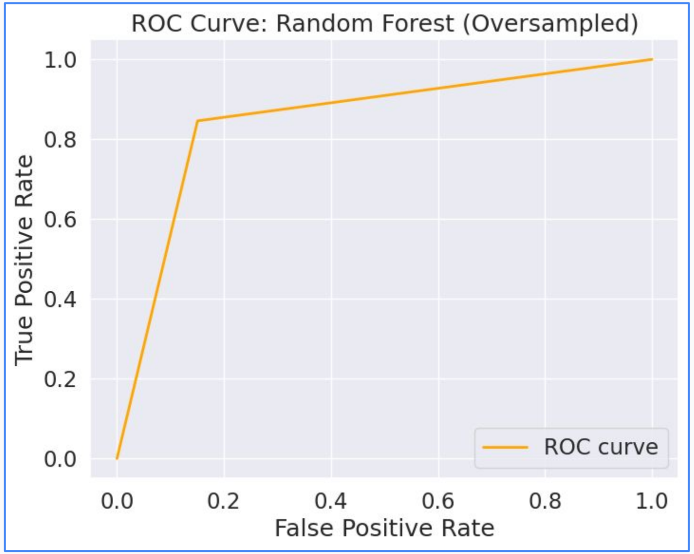


Better model ⭐

ROC Curve Plotting







Key Takeaways

Data Exploration:

- **Demographics** are **not key features** to understand customer churn.
- **Potential target audience** for improvement: **single people with no kids/dependents**.
- Customer segmentation highlights the importance of assessing customer churn.
- More than half of our customers leaving are highly profitable (spending more than average)
- **57% of customers leaving use electronic check** as a payment method (= Room for improvement)

Correlation:

- **Key features** related to customer **churn**: **contract type** and **internet service** type.
- Contract type has a moderate relationship with churn. **Focus on pitching longer-term contracts** (two years) which have the lowest churn rate.

Machine Learning Models:

- **Upsampling is crucial** for assessing machine learning models due to **significant data imbalance** (70/30 split).
- **Random Forest classifier using SMOTE** upsampling is the **best option for building an accurate model** to predict customer churn.

Thank you