

Customer Churn Prediction

***Identifying At-Risk
Customers***





Introduction

Felicia Angjaya

Finance student with a strong interest in data analytics and business insights, equipped with hands-on experience in predictive modeling and data visualization

Education

Finance, BINUS University – Semester 5



[linkedin.com/in/feliciaangjaya](https://www.linkedin.com/in/feliciaangjaya)



Felangjaya@gmail.com

Overview Project

Customer Churn Prediction : Identifying At-Risk Customers

This project focuses on predicting customer churn using historical demographic, behavioral, and service usage data. The goal is to understand what factors drive customers to leave and to build a predictive model that identifies high-risk customers before they churn. The project combines data analysis and machine learning to provide both actionable business insights and a deployable churn detection system.

Project Objectives

- 1** **Predict customer churn accurately**
Predict customer churn accurately by leveraging demographic, service usage, and behavioral features to proactively identify customers at risk of leaving
- 2** **Identify Key Drivers of Churn**
Uncover the most influential churn factors, providing actionable insights that help the business implement more targeted and effective retention strategies
- 3** **Build machine learning model**
Build and optimize a machine learning model that is both interpretable and efficient, using a simplified set of top predictive features, ensuring it can be easily deployed into real-time systems or operational workflows

Goal

The goal of this project is to proactively reduce customer churn and improve retention by leveraging machine learning to identify customers who are most likely to leave. By analyzing historical data on demographics, service usage, billing patterns, and satisfaction scores, the project aims to build a predictive model that flags high-risk customers before they churn

Tools



Streamlit

Project Background

Project Overview

This project focuses on **predicting customer churn** in a telecommunications company using historical data on customer demographics, service usage, contract types, and satisfaction. The aim is to **build a reliable machine learning model** that can accurately identify high-risk customers before they leave

Why This Project Matters ?

Customer churn poses a serious challenge to subscription-based businesses like telecom providers. Retaining existing customers is significantly more cost-effective than acquiring new ones. By understanding **who is likely to churn and why**, the business can proactively implement retention strategies, reduce revenue loss, and improve long-term customer lifetime value (CLTV)

Expected Outcomes

- Early Detection of High-Risk Customers** Accurately predict customers most likely to churn in the near future using behavior and service usage patterns
- Insight into Key Churn Drivers** Uncover the most impactful features influencing churn (e.g., contract type, satisfaction score, internet type)
- Deployable and Lightweight Model** Deliver a simplified model using top features – optimized for real-time use and seamless integration into operational systems
- Data-Driven Retention Strategies** Provide actionable insights for marketing and customer service teams to design targeted offers, campaigns, or service improvements

Beneficiaries

- 1



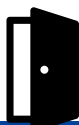
Marketing & Retention Teams :
to design personalized outreach to at-risk customers
- 2



Customer Success & Operations :
to monitor churn risk in real time and improve satisfaction
- 3



Business Strategy Team :
to align churn insights with revenue planning and resource allocation



Problem Statement

Problem Statement

1 *Difficult to Identify Who Will Churn*

It is unclear which customers are at risk of churning just by looking at raw customer data, there is no systematic way to predict churn based on their behavior and profile

2 *Need for a Data-Driven Churn Prediction Model*

A machine learning model is needed to predict churn likelihood using historical, behavioral, and demographic data

3 *No Predictive System for Early Detection*

Currently, there is no data-driven mechanism to proactively identify customers who are at risk of churning

Objectives

Predict customer churn accurately by leveraging demographic, service usage, and behavioral features to proactively identify customers at risk of leaving

Uncover the most influential churn factors, providing actionable insights that help the business implement more targeted and effective retention strategies

Build and optimize a machine learning model that is both interpretable and efficient, using a simplified set of top predictive features, ensuring it can be easily deployed into real-time systems or operational workflows

The success of this project will be measured using the following metrics :

Recall (Sensitivity)

Measures how well the model identifies actual churners, minimizing false negatives is critical in a business context

ROC-AUC Score

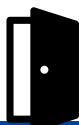
Evaluates the model's overall ability to distinguish between churners and non-churners

F1-score

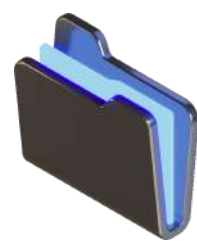
Balances precision and recall, ensuring robust performance across both metrics

Business Insight


The interpretability of the model, including feature importance and churn segmentation should provide actionable insights for strategies



Data Understanding & Preprocessing



Kaggle



Why Do Customers Leave? Can You Spot the Churners?
Dive into Customer Behavior: Can You Predict Who Stays or Goes?
kaggle.com

The dataset contains data from Telco customers in California, including demographics, service details, satisfaction scores, and churn status, intended for analyzing customer behavior and predicting churn

7043

Number of Records

53

Number of Variables

20

numerical

33

categorical

Target Variable (churn_value)

0 = tidak churn

1 = churn



No duplicate rows were found in the dataset



Note : All other columns have 0 missing values

Missing Values

| Categorical Column | Missing |
|--------------------|---------|
| churn_reason | 5174 |
| offer | 3877 |
| internet_type | 1526 |

Filled with "Unknown"

Data Transformation

numerical features

Standard Scaler

Ensures all numeric values are on the same scale (mean = 0, std = 1), improving model convergence and performance

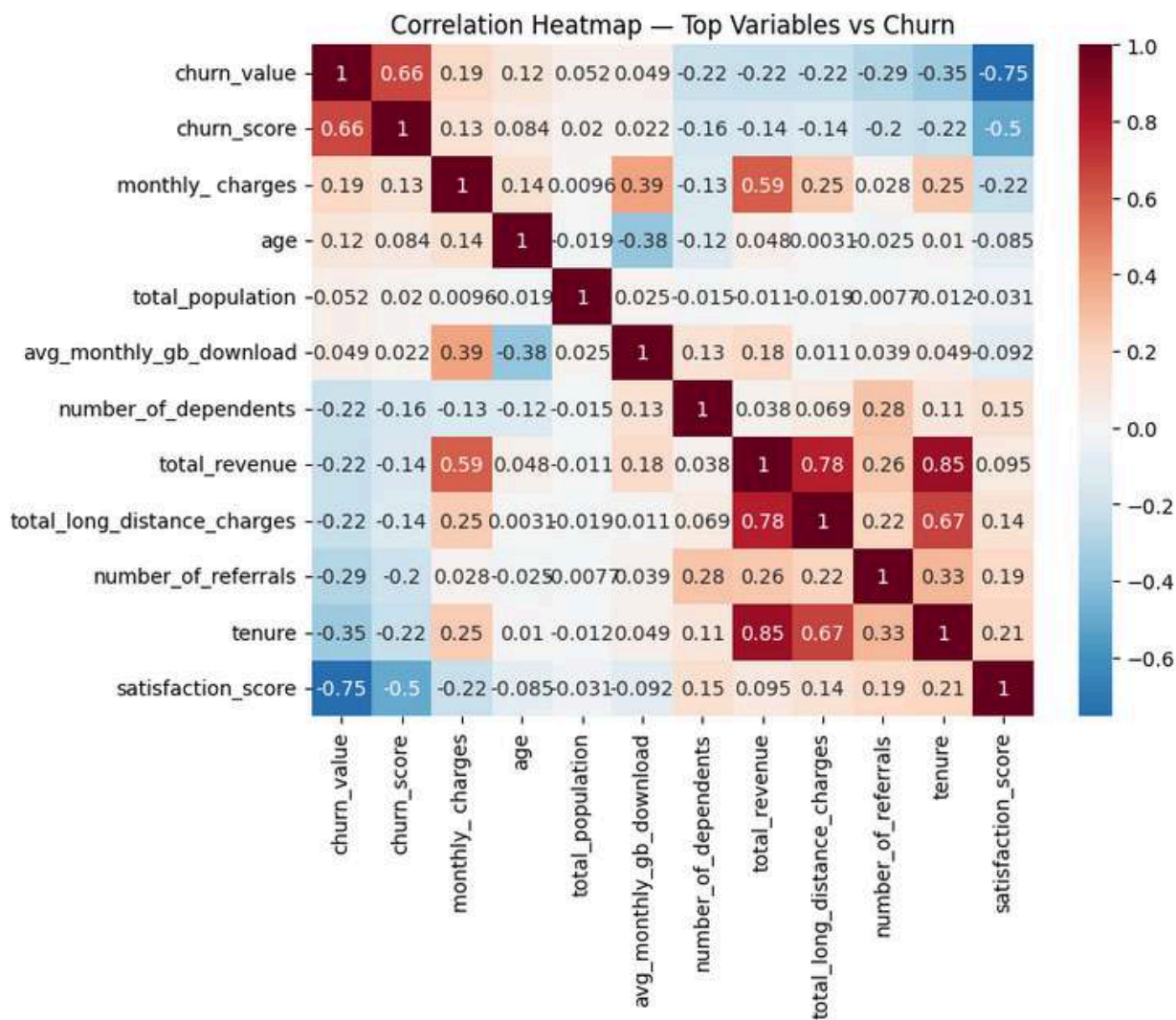
categorical features

One-Hot Encoding

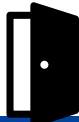
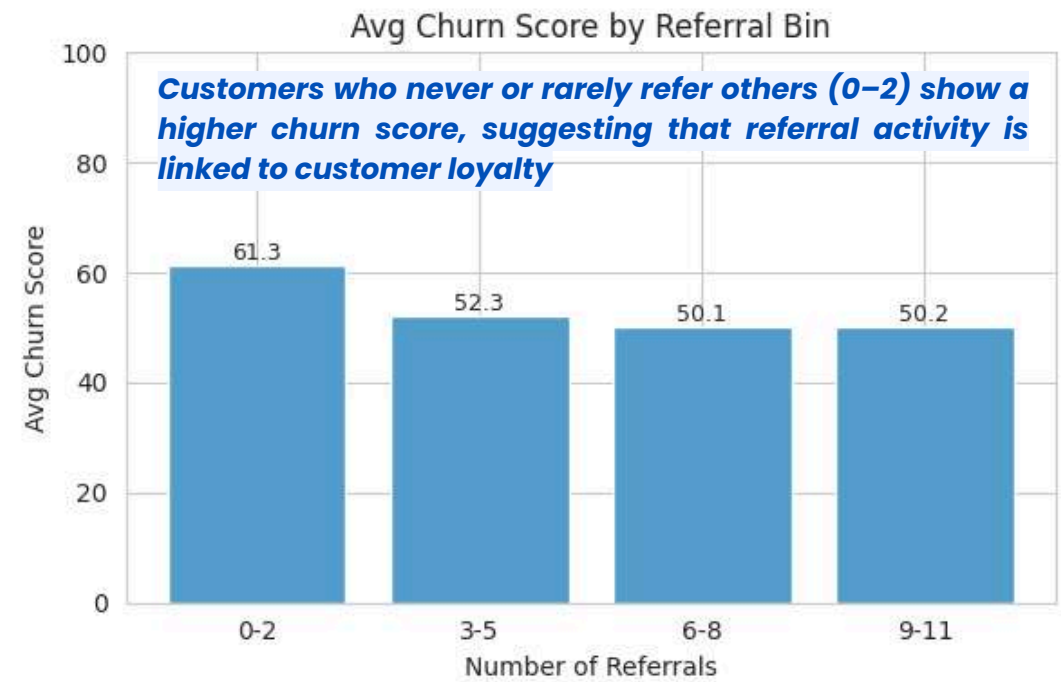
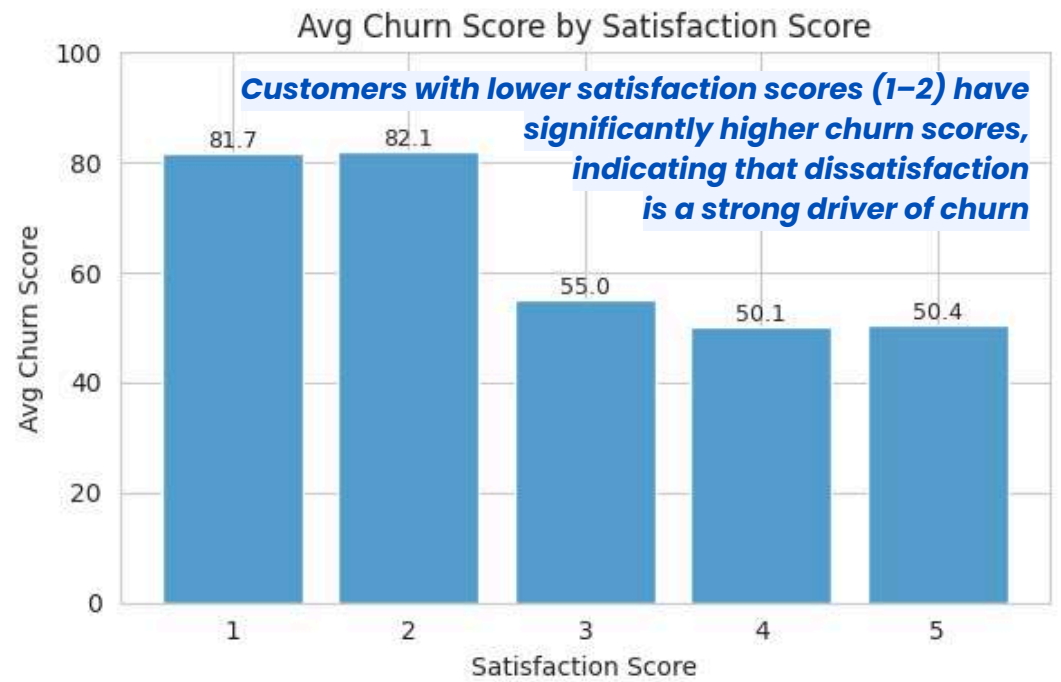
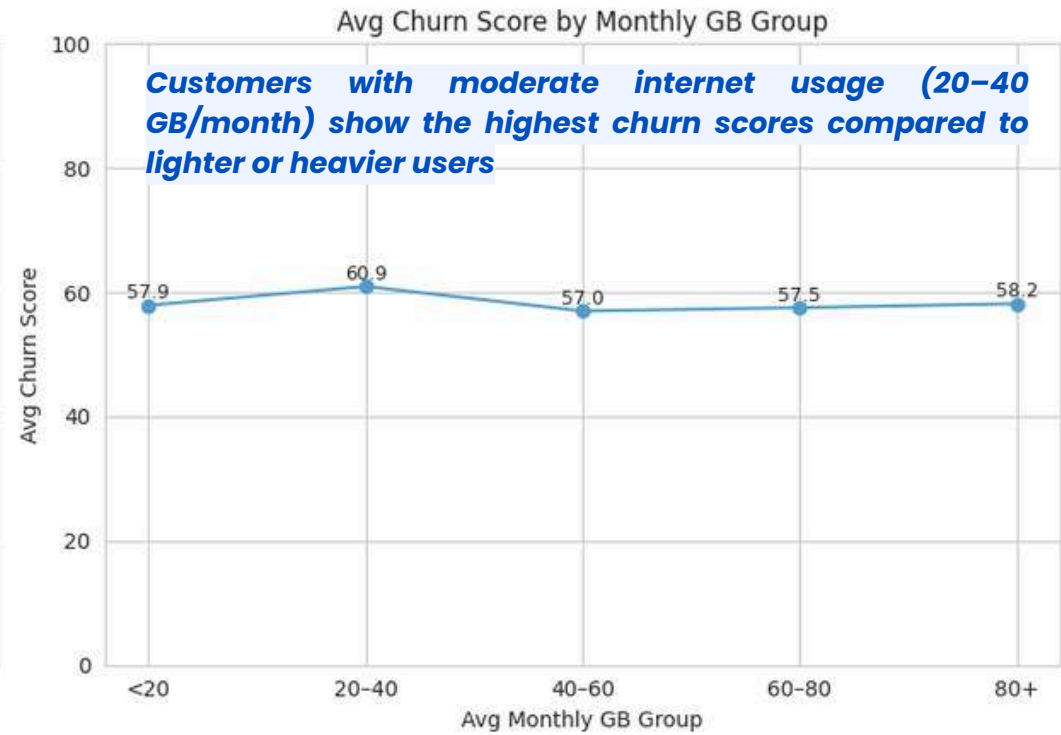
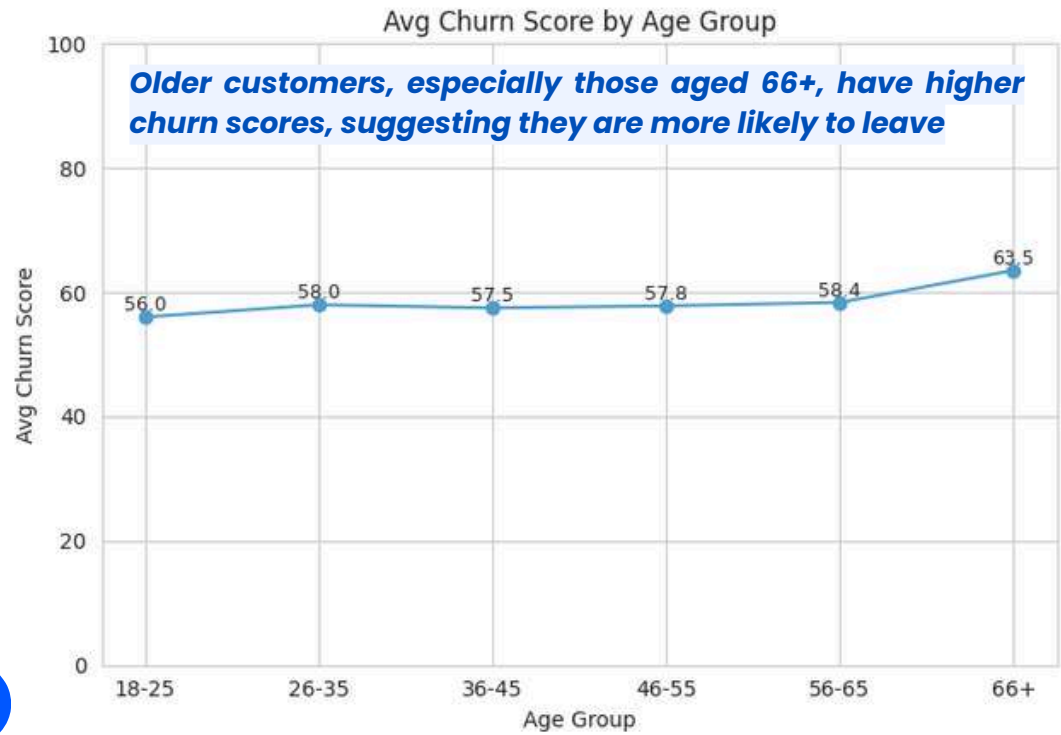
Converts categories into binary columns so that machine learning models can interpret them numerically without implying any ordinal relationship

Explanatory Data Analysis

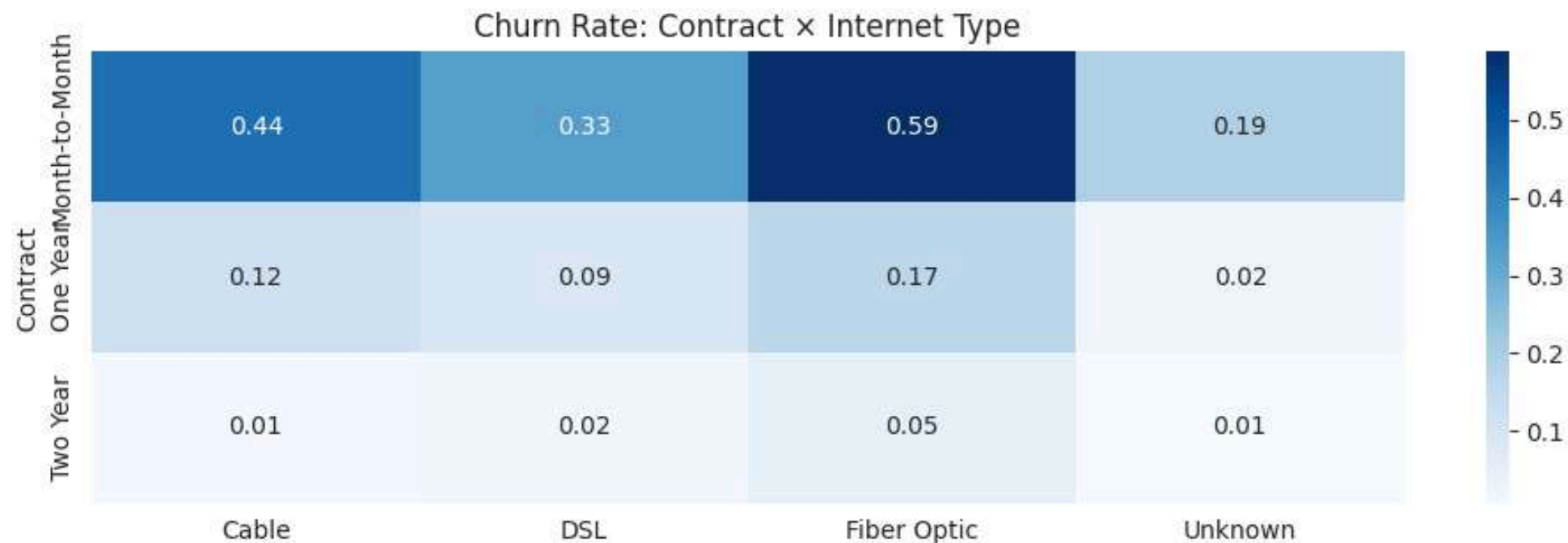
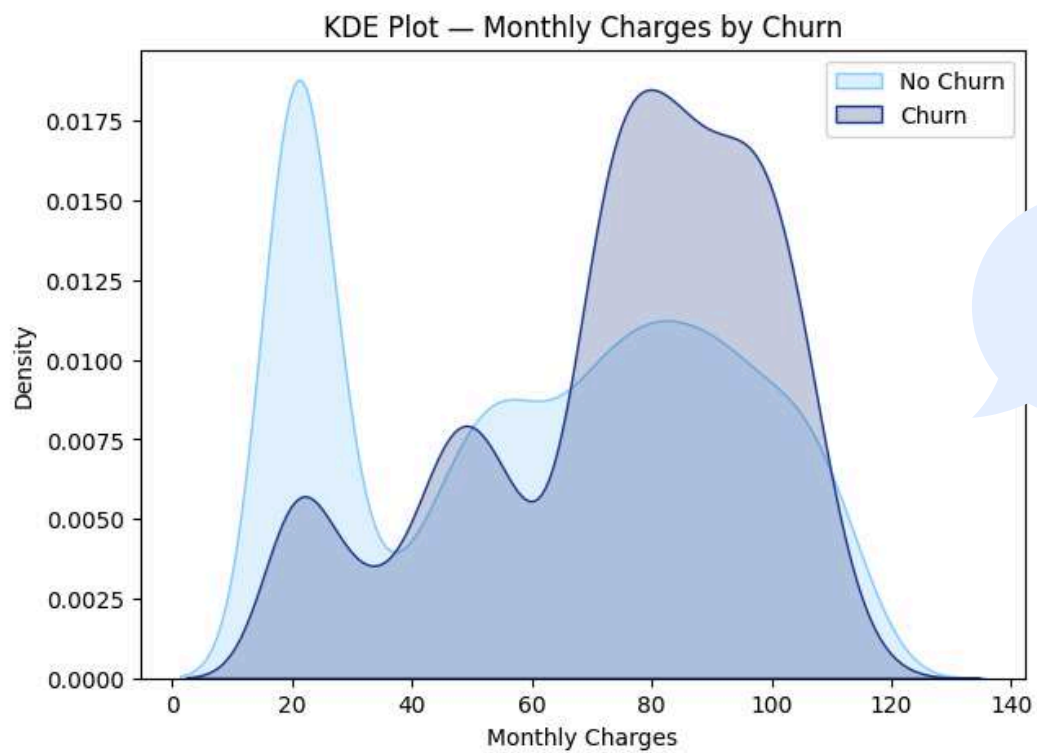
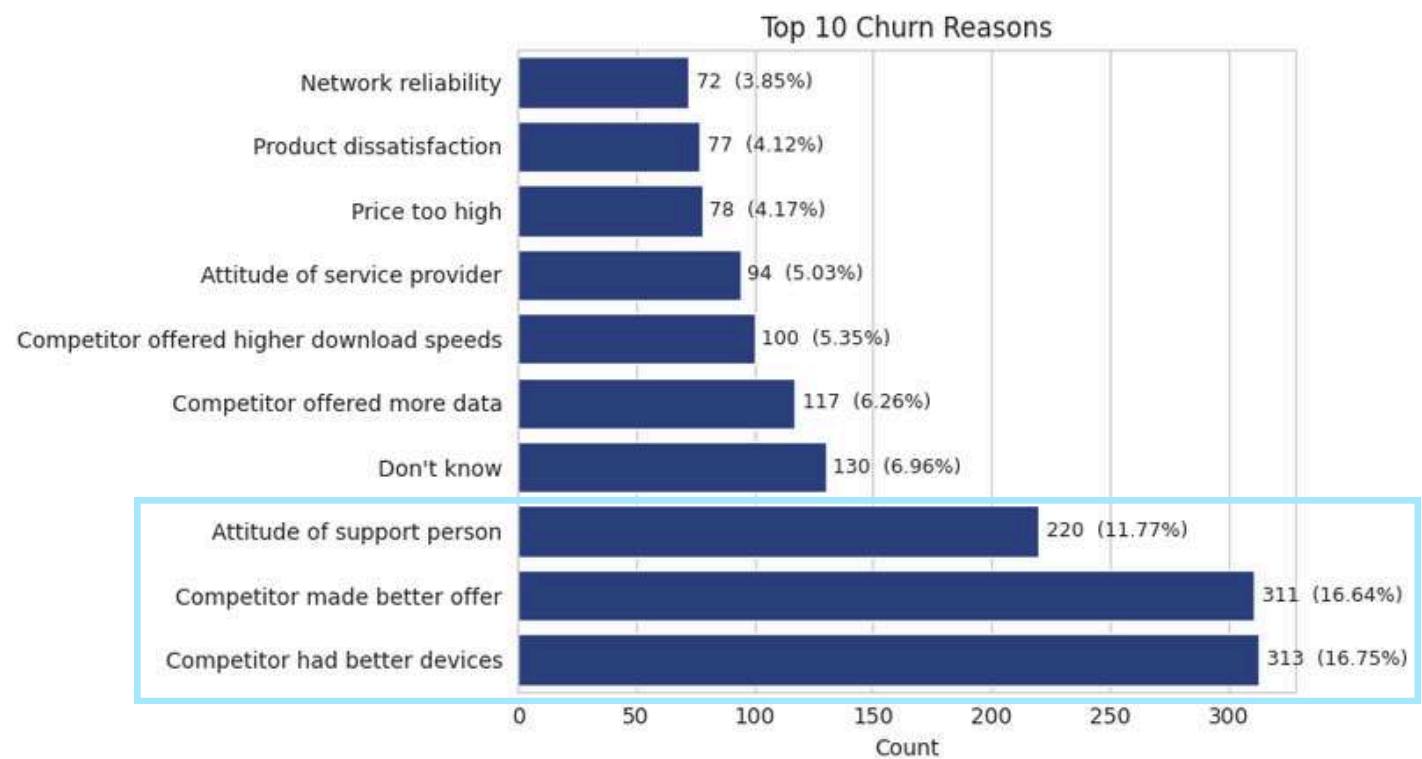
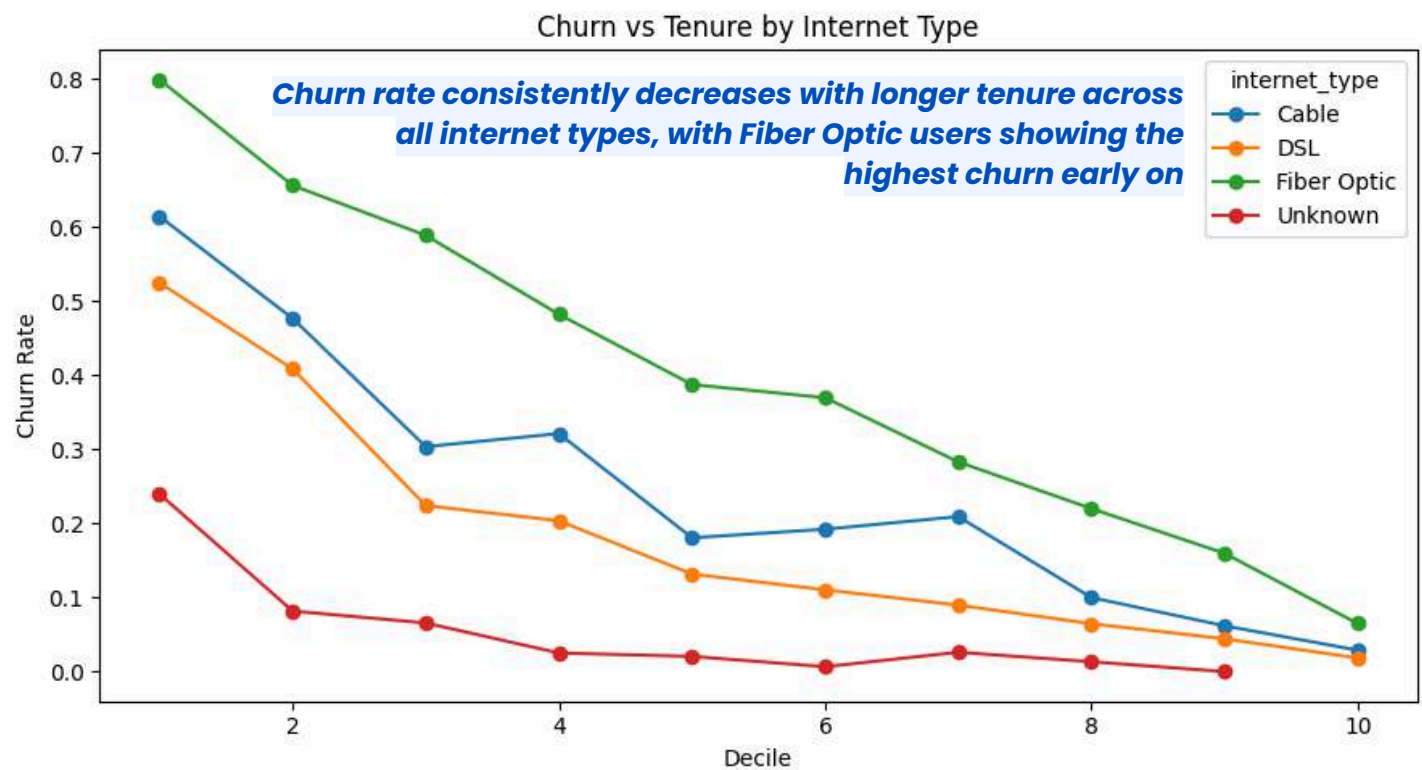
Satisfaction score shows the strongest negative correlation with churn, indicating that more satisfied customers are significantly less likely to churn. Additionally, customers with longer tenure and those who have referred others tend to be more loyal



Other factors like number of dependents, total revenue, and long-distance charges also show slight negative correlations. On the other hand, variables like age and monthly charges have very weak influence on churn likelihood.



Explanatory Data Analysis



Customers on month-to-month contracts, especially those using fiber optic internet, have the highest churn rates, while those on one or two year contracts show significantly lower churn across all internet types

Machine Learning

Data Leakage

- 1 churn_value
- 2 churn_score
- 3 churn_label
- 4 customer_status
- 5 churn_category
- 6 churn_reason
- 7 customer_id



Model Training

Trained classifiers include :

Logistic Regression

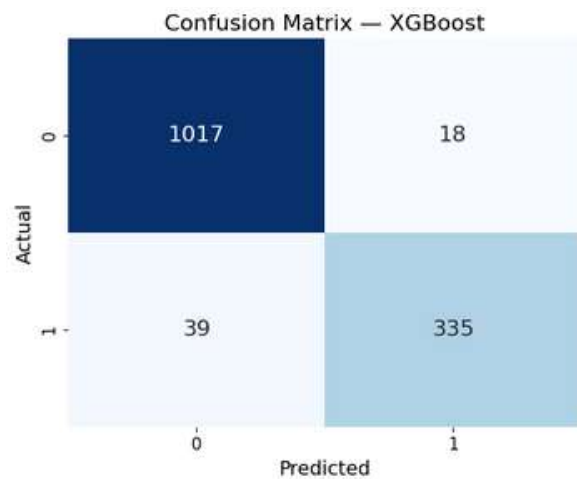
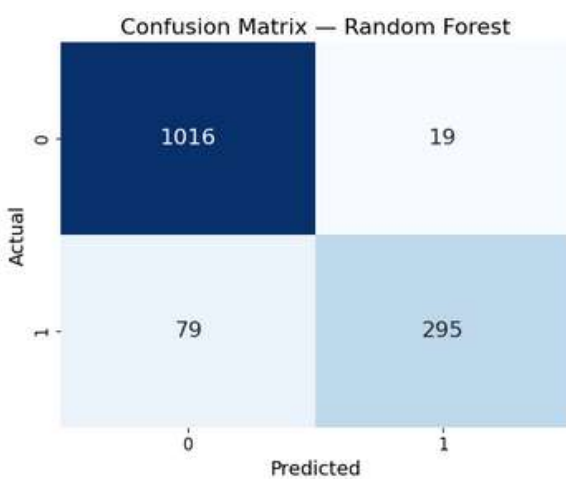
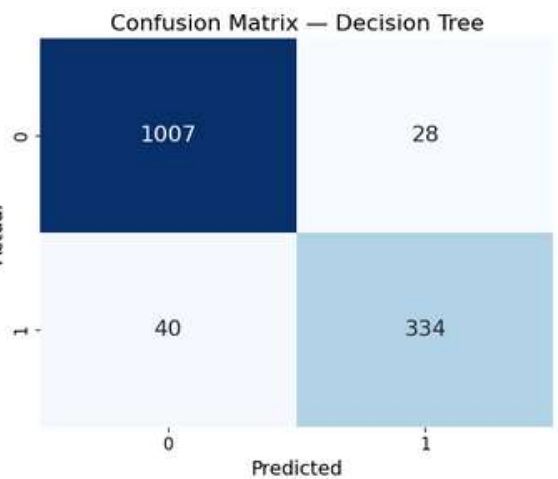
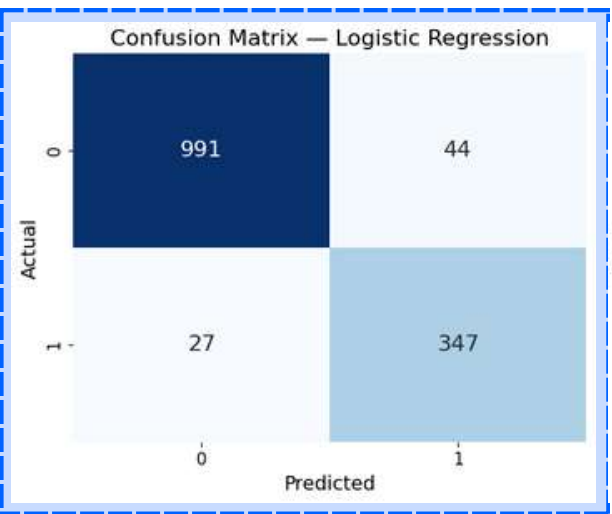
Decision Tree

Random Forest

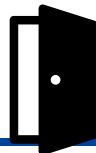
XGBoost

Model Evaluation (Evaluated using standard classification metrics)

| Model | Accuracy | Precision | Recall | F1 | ROC_AUC |
|---------------------|----------|-----------|----------|----------|----------|
| Logistic Regression | 0.949610 | 0.887468 | 0.927807 | 0.907190 | 0.990713 |
| XGBoost | 0.959546 | 0.949008 | 0.895722 | 0.921596 | 0.990700 |
| Random Forest | 0.930447 | 0.939490 | 0.788770 | 0.857558 | 0.973495 |
| Decision Tree | 0.951739 | 0.922652 | 0.893048 | 0.907609 | 0.932997 |



Chose **Logistic Regression** because it achieved the **highest recall** (92.78%), meaning it successfully detected the most actual churners among all models



Machine Learning (Final Model for Deployment)

Final Model for Deployment

To ensure the model is production-ready and efficient for real-time use, we simplified it using only the **Top-10 most important features**

Top-10 important features

city **satisfaction_score** **online_security** **dependents** **referred_a_friend**
number_of_referrals **senior_citizen** **monthly_charges** **contract** **offer**

*Coefficients from Logistic Regression were aggregated by original feature name (after one-hot encoding), and the Top-10 most impactful features were selected based on their absolute coefficient values.

Refit Model Using Top-10 Features

Simpler pipeline is built using only the Top-10 most important features

Re-trained Logistic Regression using only selected numerical and categorical features

Re-evaluated on the same test set

Logistic Regression (Top-10 Variables)

| Accuracy | Precision | Recall | F1 | ROC_AUC |
|----------|-----------|----------|----------|----------|
| 0.950319 | 0.874384 | 0.949197 | 0.910256 | 0.991397 |

Optimization Logistic Regression (Top-10 Variables)

Cross-Validation [Stratified K-Fold]

Before Tuning
CV ROC-AUC: 0.9925

Hyperparameter Tuning [GridSearchCV]

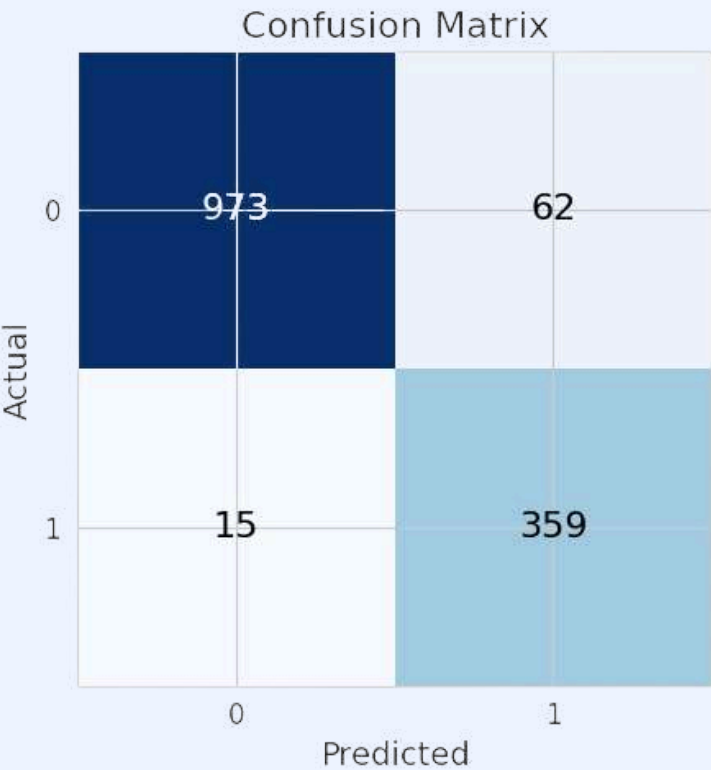
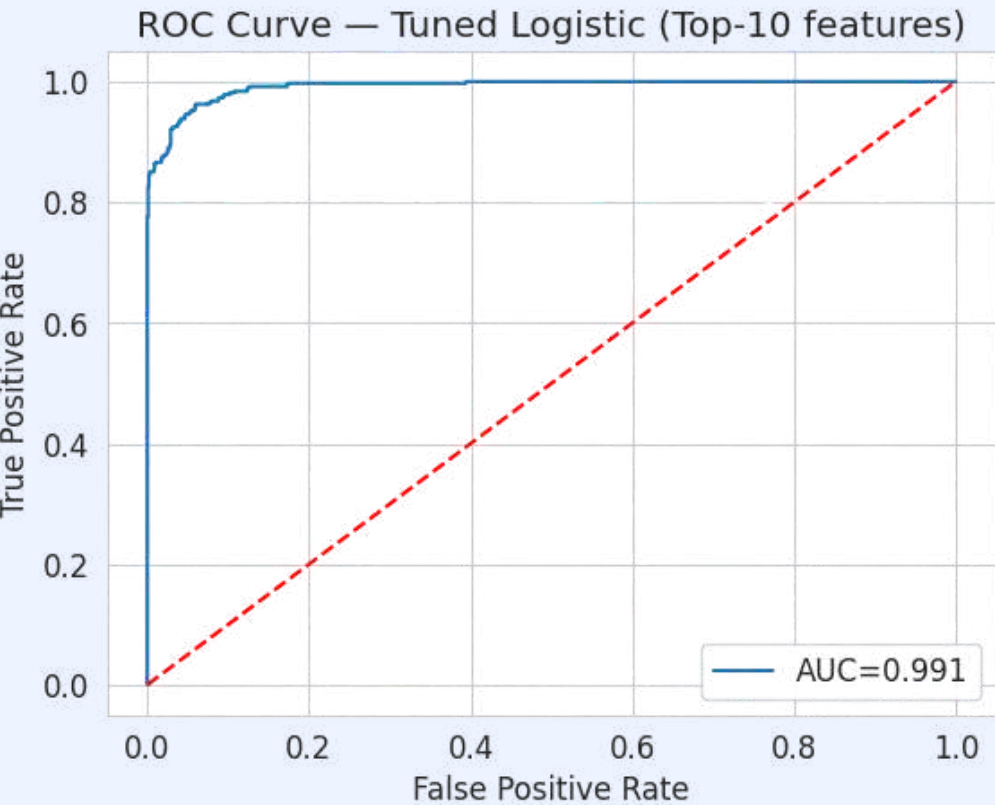
After Tuning
CV ROC-AUC: 0.9927

*Even without tuning, performance was already strong & stable

*Slight performance boost after tuning

Test Metrics

| Accuracy | Precision | Recall | F1 | ROC_AUC |
|----------|-----------|----------|----------|----------|
| 0.945351 | 0.852731 | 0.959893 | 0.903144 | 0.991477 |



Conclusion & Recommendation

Conclusion :

- The churn prediction project successfully uncovered key patterns in customer behavior that correlate with churn. The exploratory analysis showed that customers with low satisfaction, shorter tenure, and no referral activity are more prone to leaving. Additionally, higher monthly charges and month-to-month contracts, especially with fiber optic internet, were linked with higher churn rates
- Using logistic regression as the final model, chosen for its high recall score (92.78%) and interpretability, the model was able to reliably detect most churners. The model was then optimized and simplified using only the top 10 most influential features, making it suitable for business deployment and real-time monitoring

Streamlit : <http://bit.ly/4fZkswq>

Implement Targeted Retention Programs

To reduce churn effectively, retention strategies should be tailored based on the characteristics and reasons of customers most at risk. Based on churn pattern analysis and dominant churn reasons, here's a breakdown of targeted actions :

Prioritize Short-Term Contract Customers

- Offer exclusive discounts or cashback to convert them to 1- or 2-year contracts
- Provide contract extension bundles with added value (e.g. extra data, free premium support)

Address Low Satisfaction Segments

- Trigger automatic service follow-ups or apology credits for customers with low scores
- Route them to a retention specialist or VIP support line to recover the relationship

Respond to Common Churn Reasons

- Launch "Win-Back Offers" targeting customers citing competitor deals (e.g., limited-time upgrades)
- Invest in agent retraining and track churn correlation by support ticket
- Offer temporary data boosts or network optimization callbacks for speed-related complaints

Personalize Campaigns Based on Usage & Referral Data

- Build segmented campaigns offering personalized promos for low-engagement users
- Encourage referrals with double-sided incentives (e.g. "Refer & Save")



Conclusion & Recommendation

Operationalize the Predictive Churn Model

To transform the churn prediction model from analysis into business impact, it should be embedded into operational workflows. Below are the three most actionable and strategic steps for deployment :

| Integrate Churn Scores into CRM and Internal Dashboards | Launch Trigger-Based Retention Campaigns | Monitor, Evaluate, and Retrain the Model Regularly |
|---|--|---|
| Having churn risk scores directly visible in CRM or customer success tools enables teams to respond faster and more precisely | Churn prevention must be timely. Automation ensures high-risk customers receive personalized intervention before it's too late | Customer behavior and market conditions change. Without ongoing monitoring, model performance can degrade over time |

----- How to implement -----

| | | |
|---|---|--|
| Connect the trained churn model to data pipeline (weekly or monthly scoring) | Set up threshold rules (e.g., churn probability > 80%) | Track key performance metrics (e.g., recall, F1, AUC) using dashboards |
| Add churn risk scores as a field in CRM (e.g., Salesforce, HubSpot, Zoho) | Connect to your marketing tools to run A/B tests and optimize results | Gather feedback from users (e.g., "Was this alert helpful?") |
| Use visual indicators like colored tags (e.g., red = high risk) to help agents identify at-risk customers at a glance | Trigger automated actions like sending personalized offers, assigning follow-ups, or scheduling calls | Retrain the model every 3–6 months with the latest data and fine-tune the thresholds if needed |




Thank you!





Felicia Angjaya

TAB 1



Telco - Churn Prediction

This app lets you **predict churn** in a Telco Company using logistic regression model (Top-10 features)

 Manual Prediction  Batch CSV Prediction

Manual Prediction — Set the values for required features

Output variable (target): churn_value. Fill every field below, then click Predict.

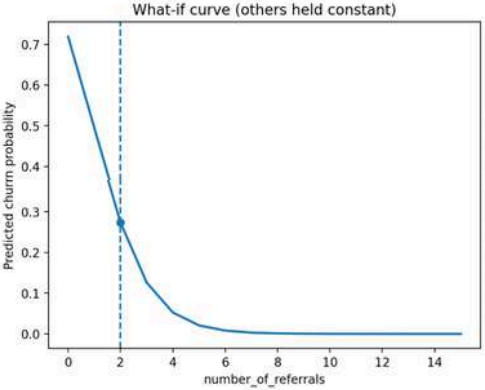
Set the inputs on the right and click **Predict**. After predicting, the left panel shows a *what-if* curve: how churn probability changes when you vary one feature (others held constant).

What-if sensitivity — Predicted vs chosen input

Pick one numeric feature; others are held constant at the values from your form.

Feature to vary

number_of_referrals



What-if curve (others held constant)

satisfaction_score

13

15

number_of_referrals

2

15

monthly_charges

65.00

-

+

city

Acampo

online_security

No

dependents

No

offer

Offer A

contract

Month-to-Month

senior_citizen

No

referred_a_friend

No


Predict

Churn probability

0.274



Predicted label (≥ 0.50): 0

TAB 2



Telco - Churn Prediction

This app lets you **predict churn** in a Telco Company using logistic regression model (Top-10 features)


 Manual Prediction  Batch CSV Prediction

Batch CSV Prediction — upload your file


Upload a CSV to score many customers at once.

Required columns:

city, satisfaction_score, online_security, dependents, number_of_referrals, offer,

 Download CSV template

Upload CSV to score

 Drag and drop file here

Limit 200MB per file • CSV

Browse files

Run batch prediction

Streamlit : <http://bit.ly/4fZkswq>