Telco Customer Churn Dataset — EDA Assignment Guide

Dataset Overview: The data pertains to the telecommunications sector (Telco) and contains 7043 rows (customers) and 21 columns.

These include demographic data (gender, senior citizen, partner, dependents),

shared services (phone service, multiple lines, internet service, cybersecurity, online backup, device protection, technical support, live TV, streaming movies),

and financial information (tenure, contract, payment method, paperless billing, monthly .(charges, total charges

Target Variable: The primary target is the Churn column, a categorical variable (Yes/No) that indicates whether the customer has left the service.

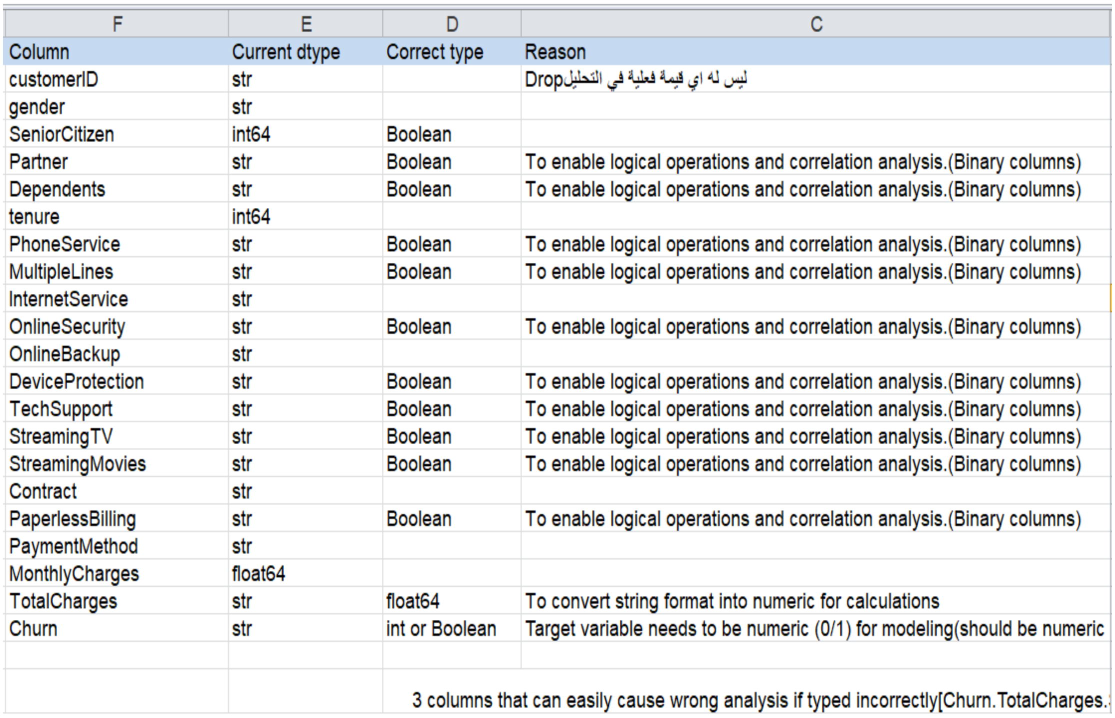
The Target Distribution: From the initial review, we observe a large number of "No" responses compared to "Yes" responses, indicating an imbalance of 73.463013% and 26.536987%

Issues Suspected

Data Type Mismatch 1/(TotalCharges): Geometrically, the TotalCharges column appears as text data or an object in some versions of this rule, whereas it should be a floating-point .number for calculations. This will break the code when attempting to calculate averages

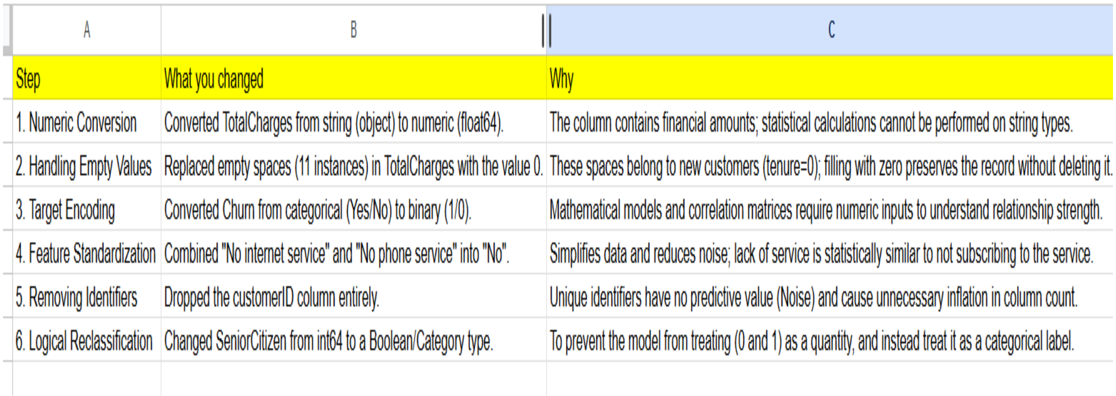
Irrelevant Features 2/(CustomerID): The customerID column is a unique identifier with no .analytical value and should be removed to reduce redundancy

Redundancy in Categories: Having values like "No internet service" in multiple columns (e.g., OnlineSecurity and TechSupport) can cause multicollinearity, requiring merging or

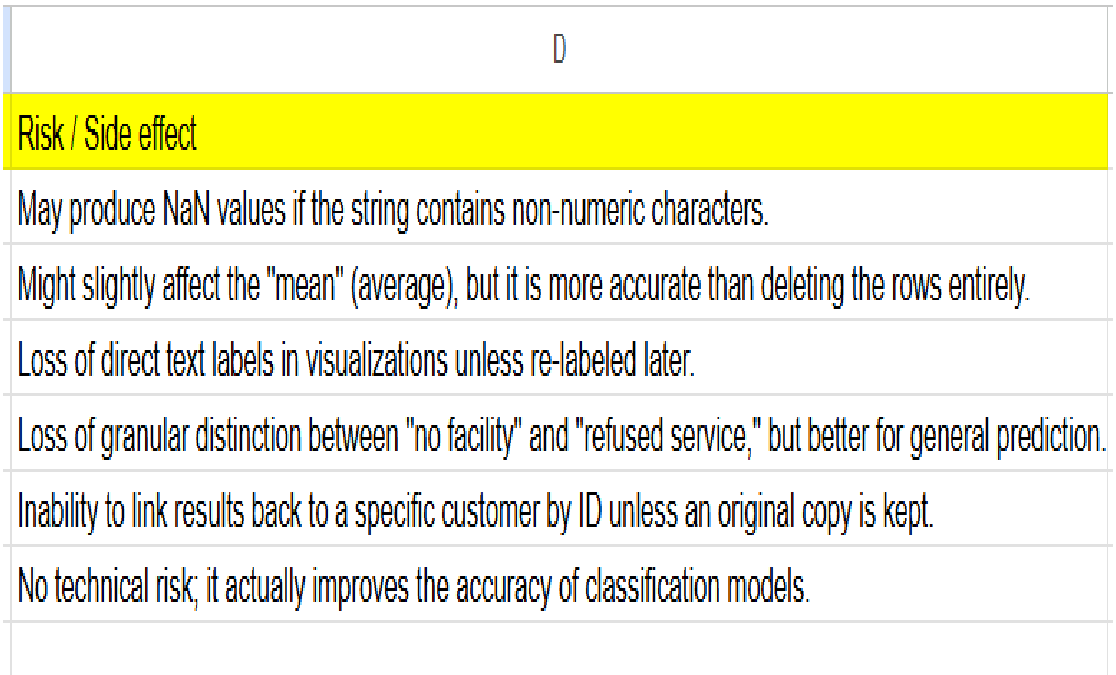


F	E	D	C
Column	Current dtype	Correct type	Reason
customerID	str		Dropليس له اي قيمة فعلية في التحليل
gender	str		
SeniorCitizen	int64	Boolean	
Partner	str	Boolean	To enable logical operations and correlation analysis.(Binary columns)
Dependents	str	Boolean	To enable logical operations and correlation analysis.(Binary columns)
tenure	int64		
PhoneService	str	Boolean	To enable logical operations and correlation analysis.(Binary columns)
MultipleLines	str	Boolean	To enable logical operations and correlation analysis.(Binary columns)
InternetService	str		
OnlineSecurity	str	Boolean	To enable logical operations and correlation analysis.(Binary columns)
OnlineBackup	str		
DeviceProtection	str	Boolean	To enable logical operations and correlation analysis.(Binary columns)
TechSupport	str	Boolean	To enable logical operations and correlation analysis.(Binary columns)
StreamingTV	str	Boolean	To enable logical operations and correlation analysis.(Binary columns)
StreamingMovies	str	Boolean	To enable logical operations and correlation analysis.(Binary columns)
Contract	str		
PaperlessBilling	str	Boolean	To enable logical operations and correlation analysis.(Binary columns)
PaymentMethod	str		
MonthlyCharges	float64		
TotalCharges	str	float64	To convert string format into numeric for calculations
Churn	str	int or Boolean	Target variable needs to be numeric (0/1) for modeling(should be numeric

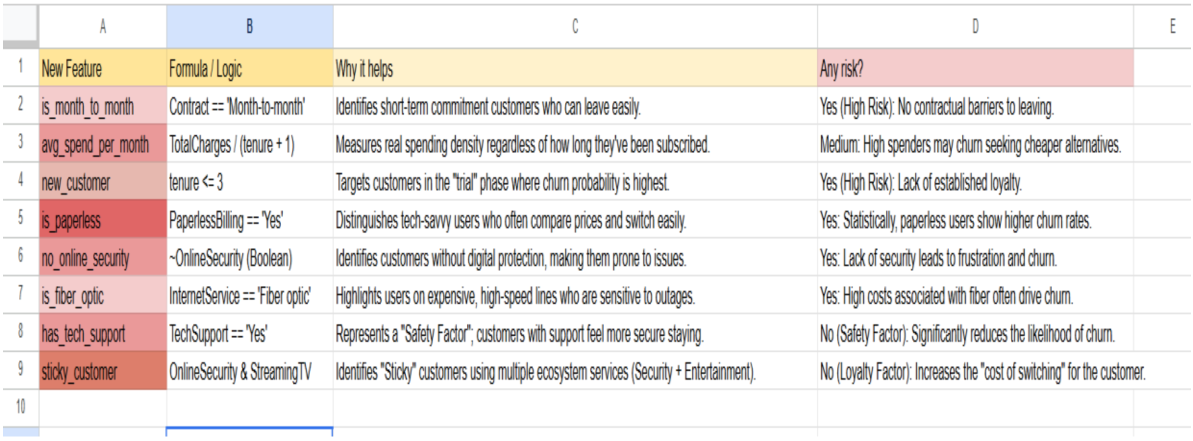
3 columns that can easily cause wrong analysis if typed incorrectly[Churn.TotalCharges.



A	B	C
Step	What you changed	Why
1. Numeric Conversion	Converted TotalCharges from string (object) to numeric (float64).	The column contains financial amounts; statistical calculations cannot be performed on string types.
2. Handling Empty Values	Replaced empty spaces (11 instances) in TotalCharges with the value 0.	These spaces belong to new customers (tenure=0); filling with zero preserves the record without deleting it.
3. Target Encoding	Converted Churn from categorical (Yes/No) to binary (1/0).	Mathematical models and correlation matrices require numeric inputs to understand relationship strength.
4. Feature Standardization	Combined "No internet service" and "No phone service" into "No".	Simplifies data and reduces noise; lack of service is statistically similar to not subscribing to the service.
5. Removing Identifiers	Dropped the customerID column entirely.	Unique identifiers have no predictive value (Noise) and cause unnecessary inflation in column count.
6. Logical Reclassification	Changed SeniorCitizen from int64 to a Boolean/Category type.	To prevent the model from treating (0 and 1) as a quantity, and instead treat it as a categorical label.



D
Risk / Side effect
May produce NaN values if the string contains non-numeric characters.
Might slightly affect the "mean" (average), but it is more accurate than deleting the rows entirely.
Loss of direct text labels in visualizations unless re-labeled later.
Loss of granular distinction between "no facility" and "refused service," but better for general prediction.
Inability to link results back to a specific customer by ID unless an original copy is kept.
No technical risk; it actually improves the accuracy of classification models.



New Feature	Formula / Logic	Why it helps	Any risk?
is_month_to_month	Contract == 'Month-to-month'	Identifies short-term commitment customers who can leave easily.	Yes (High Risk): No contractual barriers to leaving.
avg_spend_per_month	TotalCharges / (tenure + 1)	Measures real spending density regardless of how long they've been subscribed.	Medium: High spenders may churn seeking cheaper alternatives.
new_customer	tenure <= 3	Targets customers in the "trial" phase where churn probability is highest.	Yes (High Risk): Lack of established loyalty.
is_paperless	PaperlessBilling == 'Yes'	Distinguishes tech-savvy users who often compare prices and switch easily.	Yes: Statistically, paperless users show higher churn rates.
no_online_security	~OnlineSecurity (Boolean)	Identifies customers without digital protection, making them prone to issues.	Yes: Lack of security leads to frustration and churn.
is_fiber_optic	InternetService == 'Fiber optic'	Highlights users on expensive, high-speed lines who are sensitive to outages.	Yes: High costs associated with fiber often drive churn.
has_tech_support	TechSupport == 'Yes'	Represents a "Safety Factor"; customers with support feel more secure staying.	No (Safety Factor): Significantly reduces the likelihood of churn.
sticky_customer	OnlineSecurity & StreamingTV	Identifies "Sticky" customers using multiple ecosystem services (Security + Entertainment).	No (Loyalty Factor): Increases the "cost of switching" for the customer.

At the end of the process of working with and analyzing the above database in terms of data type, its deviations from the original text, and the relationships between its components, in order to predict and analyze its impact on customer attrition, I suggest adding:

/□ A counter to track the number of calls to technical support to determine the size and frequency of usage problems.

/□Offering promotions and discounts when various subscriptions increase monthly.

/□Adding customer feedback on the service.

/□Explaining the reason for choosing this location to allow for continuous improvement.

/□Increasing the number of cash payment locations in residential areas.