Telco Customer Churn Analysis
Using Machine Learning

Author: Amran Thaqif Rajendra

Tools Used: Python (Pandas, Scikit-learn, XGBoost, SHAP, Matplotlib, Seaborn), Excel, Tableau Public.

1. Introduction

Customer churn is one of the most critical challenges faced by subscription-based businesses, especially in the telecommunications industry.
It represents the percentage of customers who stop using a company's service during a given period.
For telecom providers, understanding why customers churn and how to predict it can significantly reduce revenue loss and improve customer retention.

Purpose of the Project

This project aims to analyze customer data to uncover key patterns and behaviors that contribute to churn.
By combining Python, Excel, and Tableau, the analysis provides both *data-driven predictions* and *business-level insights* to support decision-making.

Objectives

1. Identify the main factors influencing customer churn.
2. Predict which customers are most likely to churn.
3. Understand behavioral differences between loyal and churned customers.
4. Provide data-backed recommendations to reduce churn.

Key Questions

- What customer segments are at the highest risk of churn?
- How do contract type, payment method, and internet service affect churn?
- Can we build a predictive model to estimate churn probability?
- What actions can the company take to retain high-risk customers?

Tools & Technologies Used

Tool	**Purpose**
Python (Jupyter Notebook)	Data cleaning, preprocessing, EDA, machine learning
Excel (Pivot Table)	Business-level summary and simple visualization
Tableau Public	Interactive dashboards and data storytelling

2. Dataset Description

2.1 Dataset Overview

The dataset used in this project is the Telco Customer Churn Dataset from *Kaggle*, which contains demographic, service, and account information for 7,043 customers of a telecommunications company.
It serves as a realistic representation of customer behavior and churn risk factors.

Attribute	Description
customerID	Unique identifier for each customer
gender	Male or Female
SeniorCitizen	Indicates if the customer is a senior (1 = Yes, 0 = No)
Partner,Dependent	Relationship-related attributes
tenure	Number of months the customer has stayed with the company
PhoneService, MultipleLines	Telephony service details
InternetService	Type of internet service (DSL, Fiber optic, None)
OnlineSecurity, OnlineBackup, TechSupport, etc.	Add-on services
Contract	Contract type (Month-to-month, One year, Two year)
PaymentMethod	Payment method used
MonthlyCharges, TotalCharges	Billing-related variables
Churn	Target variable (Yes = customer churned, No = customer retained)

3. Data Preparation & Preprocessing

3.1 Data Cleaning

Before modeling, the following cleaning steps were executed:

- Converted *TotalCharges* from object to numeric, using pd.to_numeric(..., errors='coerce') so non-numeric entries become *NaN/NULL.* This revealed 11 nulls in *TotalCharges*.

```
df['TotalCharges'] = pd.to_numeric(df['TotalCharges'], errors='coerce')
```

```
MonthlyCharges       0
TotalCharges        11
Churn                0
dtype: int64
```

TotalCharges contains 11 Null after convert to numeric.

- Removed duplicates and trimmed unnecessary white spaces (where present).
- Ensured correct data types (MonthlyCharges, TotalCharges as numeric, tenure / SeniorCitizen as numeric/integer, categorical columns remain object).

Converting *TotalCharges* to numeric allows correct aggregation and model usage. Removing duplicates and trimming spaces reduces noise and prevents incorrect categorical splits.

3.2 Feature Engineering

New columns were introduced to support analysis and modeling:

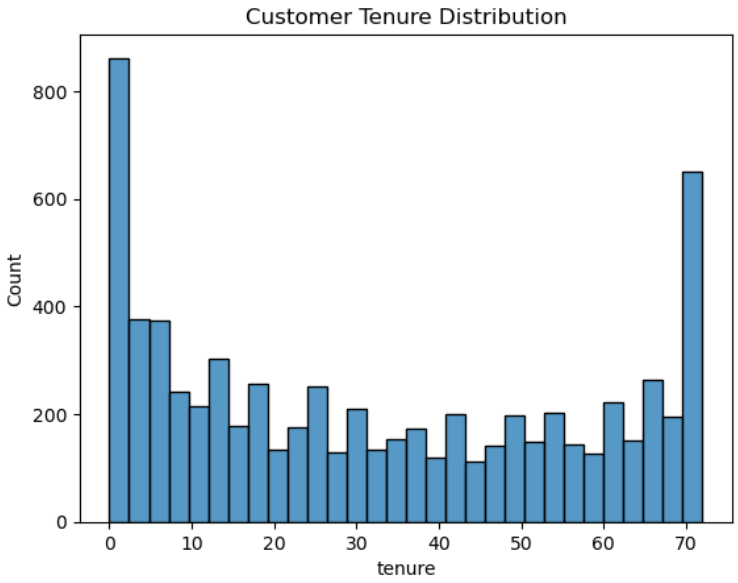
- Churn_binary: Converted Churn from Yes/No to 1/0 for numerical modeling.
- Risk-based features: Extracted or combined categorical groups (Grouped contract types).

```
df['Churn_binary'] = df['Churn'].apply(lambda x: 1 if x == 'Yes' else 0)
```

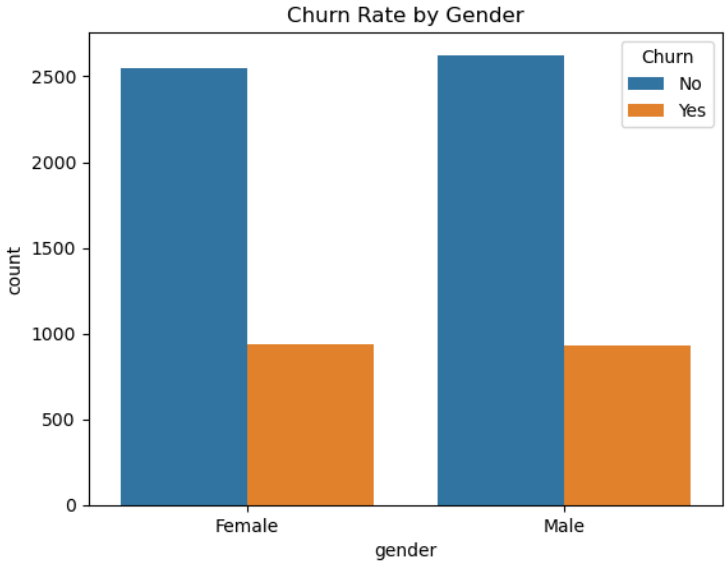
These changes ensured the dataset had both categorical and numerical variables ready for encoding.

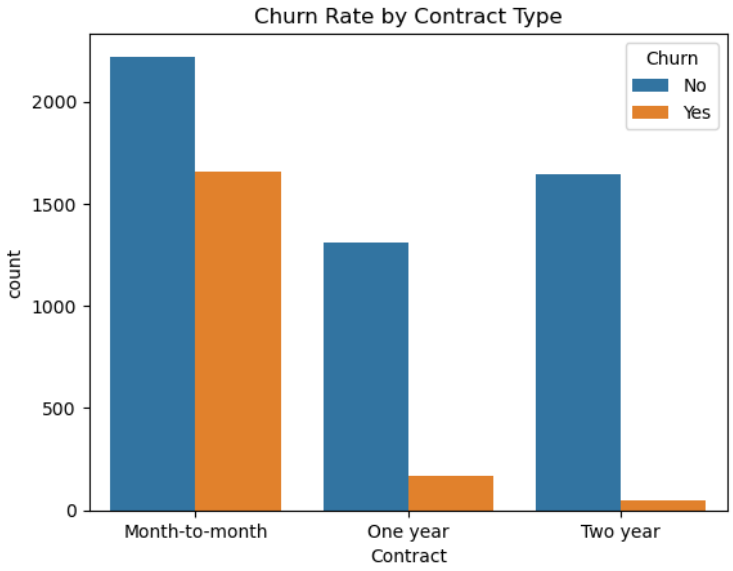
3.3 Data Visualization & Exploratory Analysis

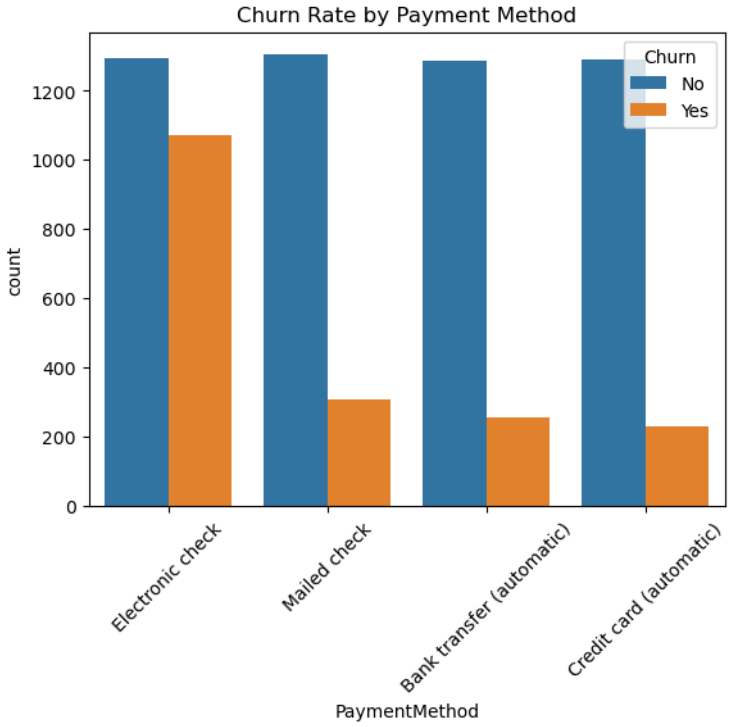
- Customer tenure distribution



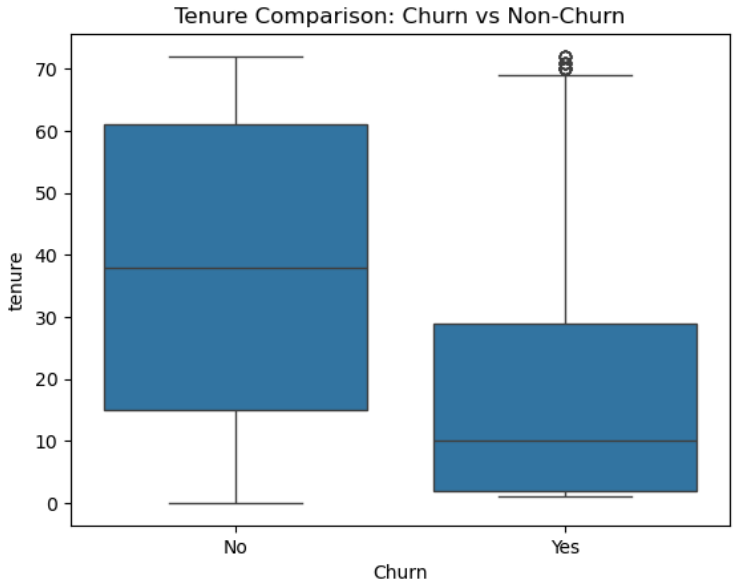
- Churn by gender, contract type, payment method




Churn Rate by Contract Type
Churn
No
Yes
2000
1500
1000
500
0
count
Month-to-month
One year
Two year
Contract


Churn Rate by Payment Method
Churn
No
Yes
1200
1000
800
600
400
200
0
count
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
PaymentMethod

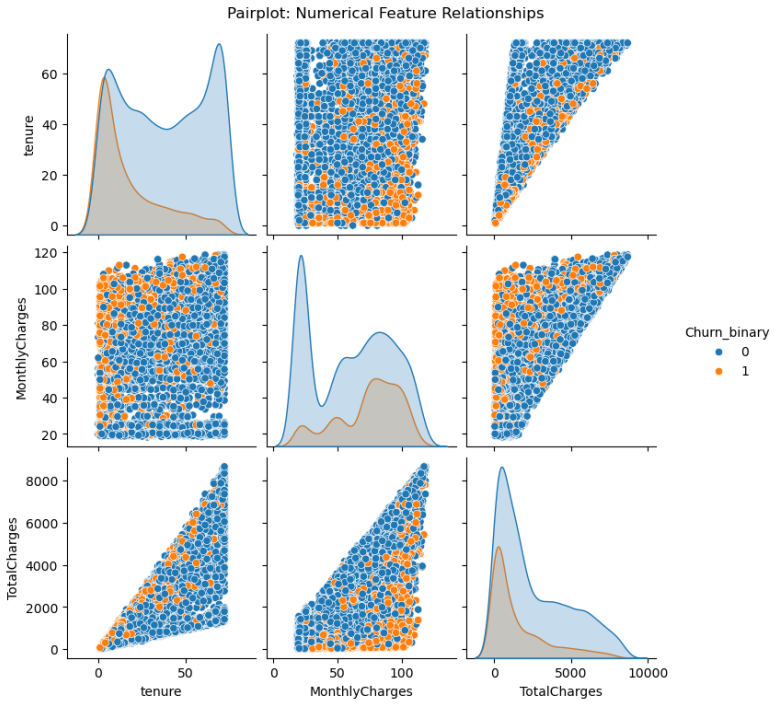
- Tenure comparison: churned vs retained



- Correlation matrix of numeric features

	tenure	MonthlyCharges	TotalCharges	Churn_binary
tenure	1.000000	0.247900	0.825880	-0.352229
MonthlyCharges	0.247900	1.000000	0.651065	0.193356
TotalCharges	0.825880	0.651065	1.000000	-0.199484
Churn_binary	-0.352229	0.193356	-0.199484	1.000000

- Pairplots to visualize relationships



Exploratory data analysis (EDA) provides business insights (e.g., month-to-month contracts churn more) and helps detect patterns for modeling.

3.4 Train-Test Split

The dataset was divided for model training and testing:

```
# ========================================
# TRAIN-TEST SPLIT
# ========================================

from sklearn.model_selection import train_test_split

# Drop unnecessary columns and define target variable
X = df.drop(['Churn', 'customerID', 'Churn_binary'], axis=1)
y = df['Churn_binary']

# Split the dataset into 80% training and 20% testing
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

- Training set: 80%
- Testing set: 20%
- Stratified split: to maintain the same churn ratio in both sets.

3.5 Data Imputation

Used median imputation for numerical features. Crucially, the median was computed from X_train and then applied to both train and test sets, this is best practice to avoid data leakage.

```
numeric_features = ['tenure', 'MonthlyCharges', 'TotalCharges']

# IMPUTATION
for col in numeric_features:
    X_train[col] = pd.to_numeric(X_train[col], errors='coerce')
    median = X_train[col].median()
    X_train[col] = X_train[col].fillna(median)

for col in numeric_features:
    X_test[col] = pd.to_numeric(X_test[col], errors='coerce')
    X_test[col] = X_test[col].fillna(median)
```

pd.to_numeric(..., errors='coerce') ensures any stray non-numeric values are safely converted to NaN before imputation. Using the training-set median prevents leaking information from the test set into the training process. This specifically addresses the 11 nulls produced when converting TotalCharges

3.6 Encoding Categorical Variables

- Categorical features were encoded using *One-Hot Encoding* to prepare them for machine learning algorithms:

```
# One-hot encode categorical variables (Change Numerical Data to Categorical Data)
X_train_encoded = pd.get_dummies(X_train, drop_first=True)

X_test_encoded = pd.get_dummies(X_test, drop_first=True)

# Align columns between train and test data
X_test_encoded = X_test_encoded.reindex(columns=X_train_encoded.columns, fill_value=0)
```

✓ 0.0s

- Ensured no missing values remain:

```
# Verify that no missing values remain

print(X_test_encoded.isnull().sum().sum())  # Output should be 0
```

✓ 0.0s

```
0
```

This step expanded categorical variables into binary dummy columns while avoiding multicollinearity by dropping the first category.

4. Predictive Modeling (Python Machine Learning)

To predict customer churn, two ensemble-based machine learning models were applied is Random Forest and XGBoost. Both algorithms are known for their strong performance on structured tabular data and their ability to handle complex feature interactions efficiently.

4.1 Model Selection Rationale

1. Random Forest (Bagging Approach)

Random Forest was chosen as a strong baseline model because it's stable, interpretable, and resistant to overfitting.
It builds multiple decision trees on random subsets of data (bagging) and averages their predictions to improve overall accuracy.

This approach works well for the Telco Churn dataset, which contains both categorical and numerical variables. It's also relatively robust to outliers and missing values.

2. XGBoost (Boosting Approach)

XGBoost (Extreme Gradient Boosting) was used as a performance-oriented model. Unlike Random Forest, it builds trees sequentially, where each new tree focuses on correcting the errors made by the previous ones. This boosting mechanism allows XGBoost to capture non-linear and complex relationships more effectively, making it one of the most widely used algorithms in churn prediction.

Why Only Two Models Were Used

The purpose of this analysis was to focus on models that are powerful, efficient, and well-suited for churn prediction without overcomplicating the workflow. Other algorithms such as Logistic Regression, SVM, or KNN were excluded because they typically require extensive preprocessing (feature scaling, encoding, etc.) and are less effective in handling non-linear feature relationships found in customer churn datasets.
By using Random Forest and XGBoost, the analysis maintains a balance between interpretability, accuracy, and computational efficiency.

Why Hyperparameter Tuning Was Not Applied

Both models were trained using default parameters to establish a baseline performance. The aim of this stage was to demonstrate the end-to-end analytical process, from data preparation to model evaluation, rather than optimizing for the highest possible

accuracy.
Hyperparameter tuning (GridSearchCV or RandomizedSearchCV) can be implemented in future stages to further refine model performance once the baseline is established.

4.2 Modeling Approach

- Models used: Random Forest and XGBoost classifiers
- Goal: Predict churn probability (Churn_Probability) for each customer
- Data: Preprocessed dataset with numeric features, categorical variables one-hot encoded, missing values imputed, and train-test split applied.

Random Forest is robust for general predictions, while XGBoost is more powerful for imbalanced datasets, common in churn problems.

4.3 Model Perfomance

Random Forest Result

Metric	**Score**
Accuracy	0.80
Precision	0.79
Recall	0.80
F1-Score	0.79
AUC-ROC	0.821

XGBoost Result

Metric	**Score**
Accuracy	0.76
Precision	0.79
Recall	0.76
F1-Score	0.77

AUC-ROC	0.832

Random Forest slightly better overall accuracy, but XGBoost achieves higher AUC-ROC, indicating better ability to distinguish churners vs non-churners. Focus is on precision and recall for churned customers (minority class).

```
AUC-ROC (RandomForest): 0.8211359115451186
              precision    recall  f1-score   support

           0       0.83      0.90      0.87      1035
           1       0.65      0.51      0.57       374

    accuracy                           0.80      1409
   macro avg       0.74      0.70      0.72      1409
weighted avg       0.79      0.80      0.79      1409
```

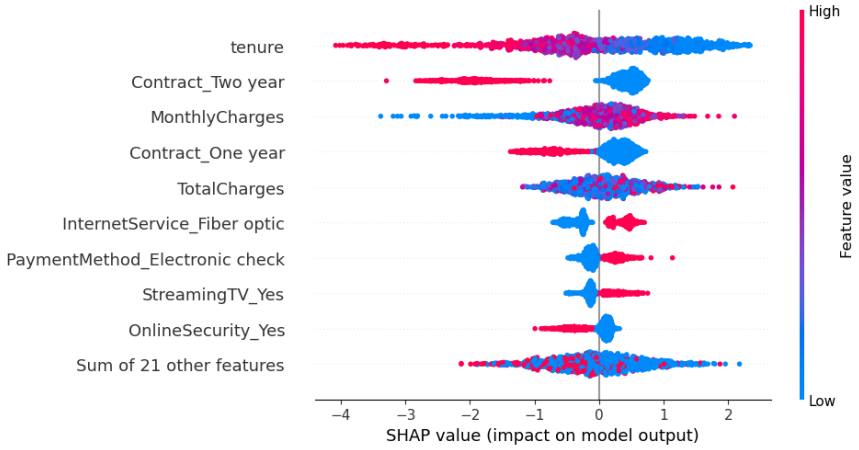
```
  AUC-ROC (XGBoost): 0.8315675940995633
               precision    recall  f1-score   support

            0       0.88      0.78      0.83      1035
            1       0.54      0.70      0.61       374

     accuracy                           0.76      1409
    macro avg       0.71      0.74      0.72      1409
 weighted avg       0.79      0.76      0.77      1409
```

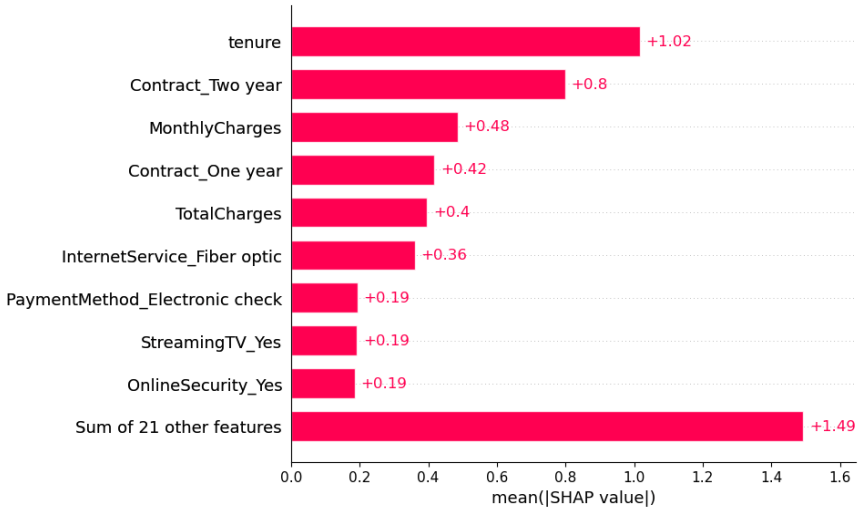
4.4 Feature Importance (SHAP Analysis)

SHAP (SHapley Additive exPlanations) was used for interpreting model predictions. Top features influencing churn:

1. Contract type, Month-to-month contracts increase churn risk.
2. Payment method, Electronic check users more likely to churn.
3. Monthly charges, Higher bills correlate with higher churn probability.
4. Tenure, New customers have higher churn.
5. Internet service type, Fiber optic users show higher churn.

SHAP allows both technical evaluation (quantifying feature contribution) and business insight (understanding why customers leave).





4.5 Prediction Results

Predicted churn probability and classification added back to the dataset for all customers.

4.6 Risk level classification:

Churn Probability	Risk Level
0-0.5	Low Risk
0.5 - 0.75	Medium Risk
0.75 - 1	High Risk

Example Output :

customerID	Churn	Predicted_Churn	Churn_Probability	Risk_Level
4376-KFVRS	No	0	0.001329	Low Risk
2754-SDJRD	No	1	0.977930	High Risk

Business teams can prioritize retention campaigns for Medium & High Risk customers.

5. Exploratory Data Analysis (Excel Pivot Analysis)

Excel is used here to provide a business-friendly exploration of customer churn. Pivot tables and charts help visualize patterns without needing technical knowledge of Python or machine learning.

5.1 Added New Loyalty Feature

A new categorical column named “Loyalty” was created in Excel to classify customers based on their tenure duration, using the formula :

```
=IF(F2<=12; "New"; IF(F2<=36; "Mid"; "Loyal"))
```

This segmentation divides customers into :

- New : Tenure < 6 months
- Medium : Tenure between 6 and 24 months
- Loyal : Tenure > 24 months

This feature helps identify that new customers are more likely to churn, while loyal customers tend to stay longer, providing a clear behavioral insight for retention strategies.

5.2 Churn Rate Overview

- Using a pivot table, we calculated the overall churn rate.

- Insight: About 27% of customers churned.
- Business interpretation: Roughly 1 in 4 customers are leaving, which is significant for revenue planning.

Row Labels	Count of customerID
No	5174
Yes	1869
Grand Total	**7043**

Churn Rate	27%

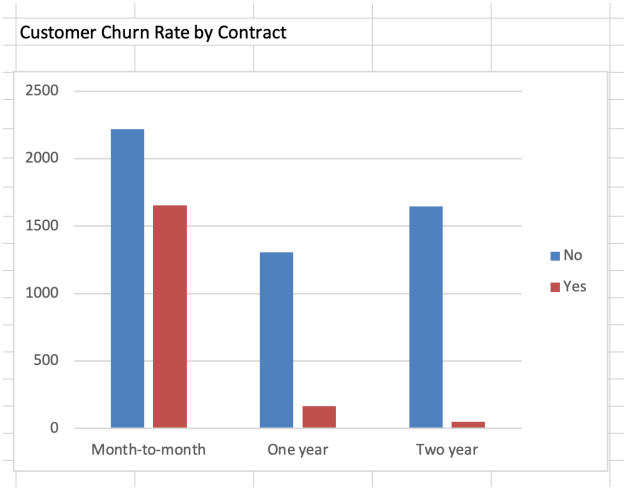
Row Labels	Average of Churn_Binary
Loyal	12%
Mid	26%
New	47%
Grand Total	**27%**

This gives a high-level view of retention risk. Business managers can quickly understand what proportion of customers are leaving.

5.3 Churn by Contract Type

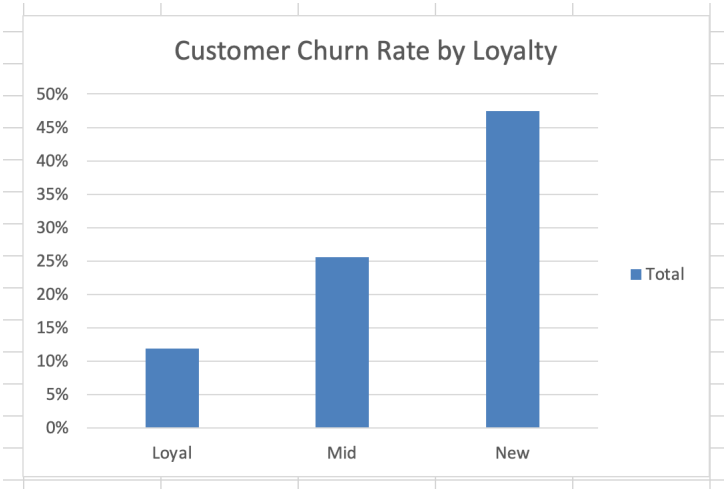
- Pivot table grouped customers by Contract type: Month-to-month, One year, Two year.
- Finding: Month-to-month contracts have the highest churn rate.

- Business interpretation: Short-term contracts increase customer turnover; longer contracts improve retention.



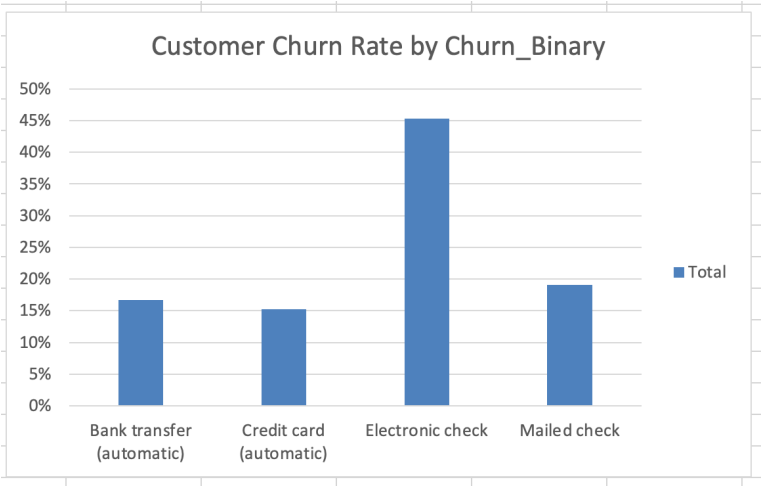
5.4 Churn by Loyalty (Tenure)

- Pivot table categorized customer tenure into groups (e.g., 0-12 months, 13-36 months, etc.).
- Finding: New customers with shorter tenure are more likely to churn.
- Business insight: Early engagement and retention programs are crucial in the first year.



5.5 Churn by Payment Method

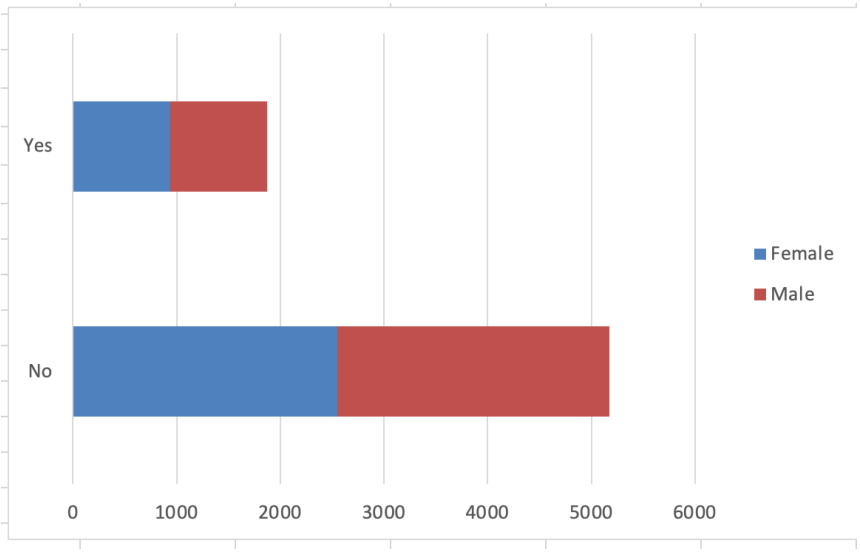
- Pivot table analyzed churn by Payment Method: Electronic check, Mailed check, Bank transfer, Credit card.
- Finding: Customers paying via Electronic Check have the highest churn.
- Business insight: Payment convenience impacts retention, offering easier payment options could reduce churn.



5.6 Churn by Gender

- Pivot table compared churn across gender.
- Finding: Minimal difference between male and female churn.

- Business insight: Gender alone is not a significant factor, focus should be on contract type, payment method, and tenure.



Pivot tables and charts allow managers to see and understand churn patterns quickly without technical background. All pivot analyses use the same cleaned dataset as Python modeling, ensuring consistency.

6. Visualization & Dashboarding (Tableau)

To transform predictive and analytical results into an interactive visual format that allows business users to explore customer churn patterns clearly. Tableau dashboards help stakeholders quickly understand which factors contribute to churn and identify at-risk customer groups.

6.1 Dashboard 1: Overall Overview

Scatterplot Monthly Charges vs Churn Probability

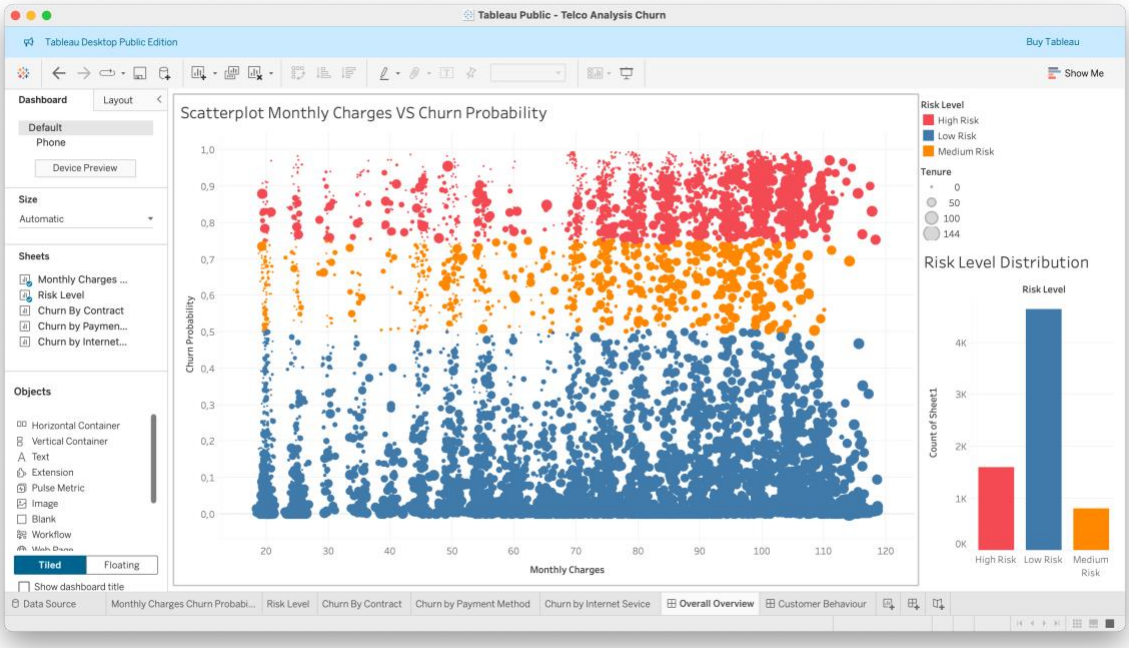
- Description: This scatter plot shows the relationship between customers *Monthly Charges* and their *Churn Probability*.

- Insight: Each point represents a customer. A clear upward trend suggests that higher monthly charges are associated with a higher probability of churn, indicating that pricing plays a major role in customer decisions.

Risk Level Distribution

- Description: This bar chart illustrates the distribution of customers across *Low*, *Medium*, and *High* churn risk levels.
- Insight: The majority of customers fall under the *Low Risk* category, but the *High Risk* segment remains significant enough to justify targeted retention strategies.

Dashboard Visualization:



This dashboard provides a strategic overview of churn likelihood and customer segmentation. Business teams can use it to prioritize which customer groups require intervention, for instance, offering discounts or personalized service to high-churn clusters.

6.2 Dashboard 2: Customer Behaviour

Churn by Contract Type

- Description: Shows how customer churn varies across contract types.
- Insight: The highest churn occurs among *Month-to-Month* contract users, while *Two-Year* contract customers show much lower churn, indicating that longer-term commitments help reduce churn.

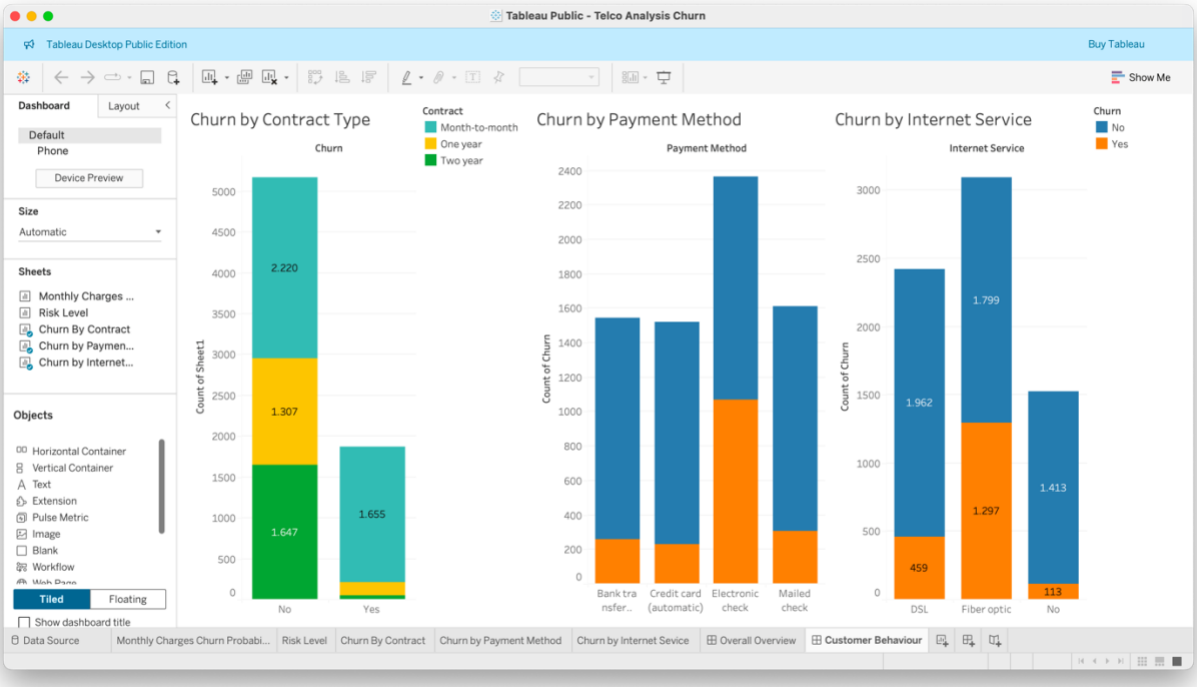
Churn by Payment Method

- Description: Displays churn distribution by payment method.
- Insight: *Electronic check* users churn significantly more than those paying via automatic credit card or bank transfer, likely due to convenience or billing issues.

Churn by Internet Service

- Description: Visualizes churn across internet service types.
- Insight: *Fiber Optic* users have the highest churn rate, possibly due to higher costs or dissatisfaction, while *No Internet Service* customers churn the least.

Dashboard Visualization:



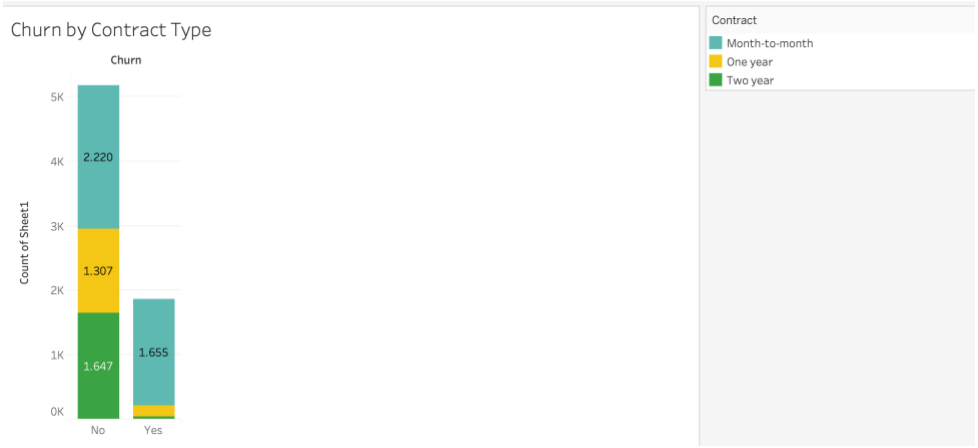
This dashboard focuses on customer behaviour patterns, helping identify service types, contract terms, and payment methods that correlate with churn. It enables the business to tailor loyalty programs and improve retention in the highest-risk segments

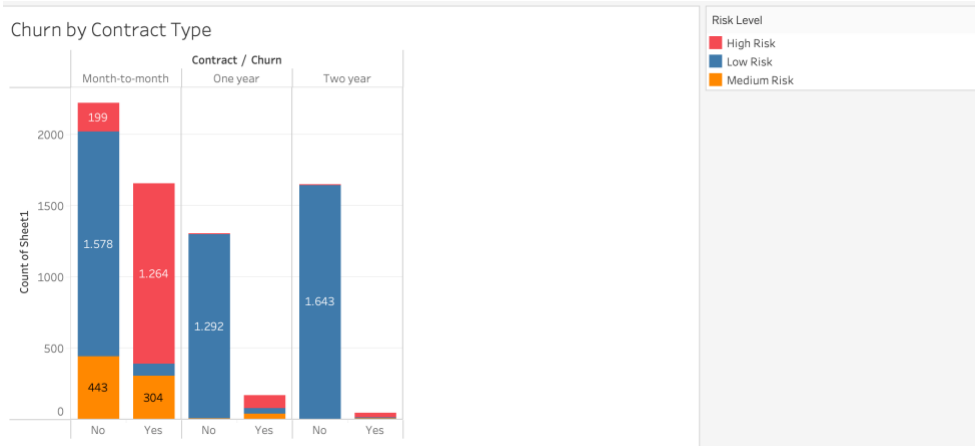
7. Insights & Business Recommendations

To summarize the key findings from the Telco churn analysis (Python, Excel, Tableau) and provide actionable recommendations that a telecom company can implement to reduce churn.

7.1 Key Insights

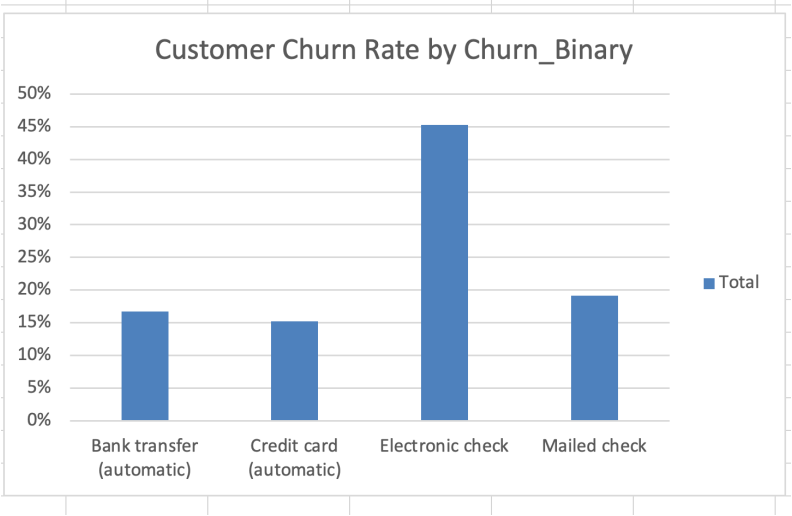
1. **Contract type drives churn**
   Month-to-month customers are most at risk of churn. Shorter contracts create uncertainty, longer commitments improve retention.
   Tableau Visualization :





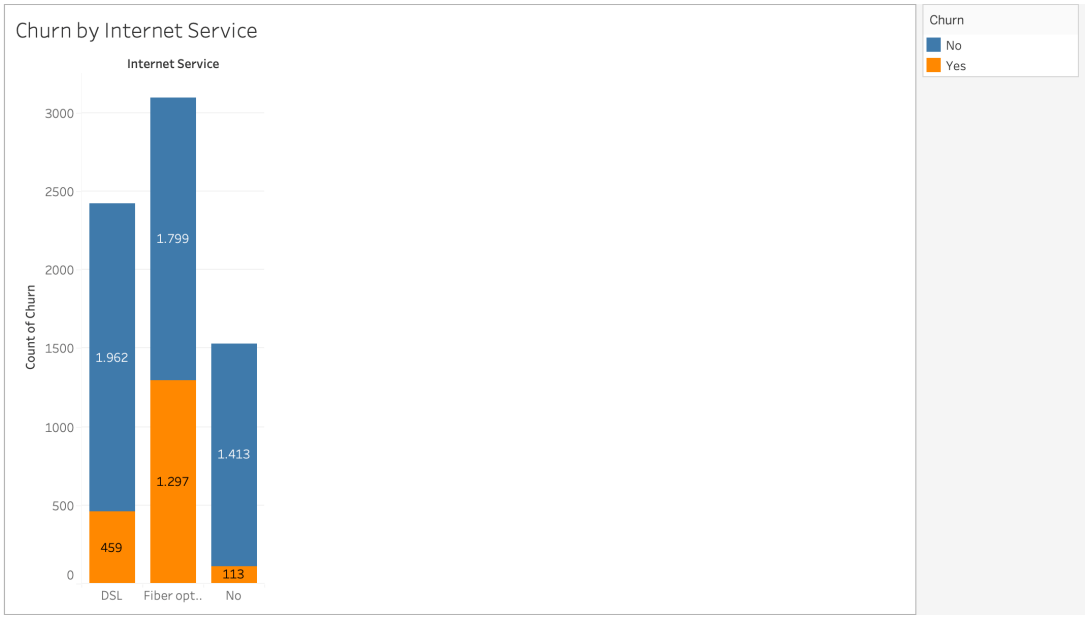
2. **Payment method affects retention**

   Customers using Electronic Check have the highest churn. Payment convenience matters, friction in payment processing increases churn probability.



3. **Internet service type impacts churn**

   Fiber optic users churn more than DSL or no internet customers. Possibly due to higher monthly charges or service dissatisfaction.



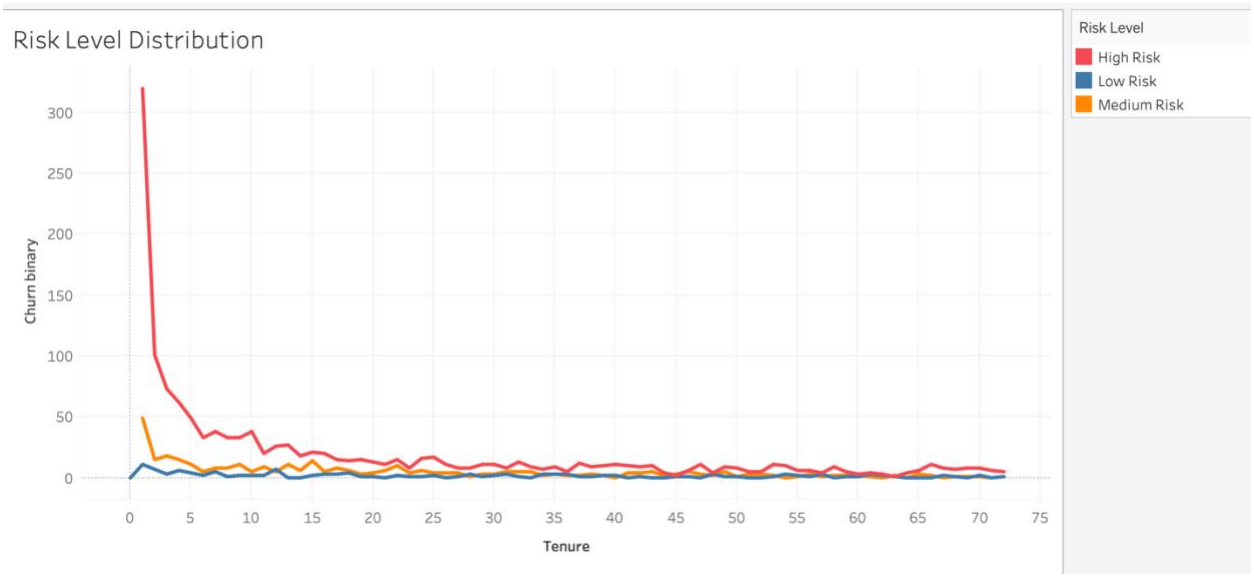
4. **Monthly charges correlate with churn probability**

   Higher monthly charges are associated with higher churn. Cost-sensitive customers are more likely to leave if perceived value is low.

5. **Customer tenure is a predictor of churn**

   New customers churn more than loyal customers. Early retention programs are critical in the first months.

   Tableau Visualization :



7.2 Business Recommendations

1. **Encourage longer contracts**

   Offer incentives like discounts, bundled services, or loyalty programs to move customers from month-to-month to one- or two-year contracts.

2. **Simplify payment options**

   Promote auto-payment methods like credit card or bank transfer to reduce churn among electronic check users.

3. **Optimize pricing and service for fiber optic users**

   Review pricing strategy and customer experience to ensure high-value services justify the cost.

4. **Target high-risk segments with retention campaigns**

   Use churn probability and risk level classifications from Python XGBoost

predictions to identify and proactively retain at-risk customers.

5. **Focus on early retention programs**
   Implement onboarding campaigns, personalized offers, and check-ins during the first 3-6 months of customer tenure.

8. Conclusion

8.1 Summary of the Project

This Telco Churn Analysis project demonstrates an end-to-end workflow integrating Python, Excel, and Tableau to understand customer churn and provide actionable insights.

8.2 Key Takeaways

1. **Python (Technical Analysis)**

- Performed data cleaning, preprocessing, feature encoding, and predictive modeling using Random Forest and XGBoost.
- Generated churn probabilities and risk level classifications for all customers.
- Identified key predictive features such as contract type, payment method, monthly charges, and tenure.

2. **Excel (Descriptive & Exploratory)**

- Created pivot tables to explore churn across key segments: contract type, payment method, gender, and tenure.
- Added a new categorical variable named “Loyalty” to represent customer loyalty segments based on tenure duration.
- Provided easy-to-understand visualizations for business stakeholders.

3. **Tableau (Interactive Dashboards)**

- Built Overall Overview and Customer Behaviour dashboards.

- Visualized patterns such as high churn among month-to-month contracts, fiber optic users, and electronic check payers.
- Enabled interactive exploration for strategic decision-making.