Telco Customer Churn Analysis & Prediction

Team: CLS ONL3_AIS4_G1 – Team 2

Team Members:

1- Shahd Ashraf Ramadan

2- Aya Anwar Mahmoud

3- Aya Hesham Saleh

4- Ayat Mohammad Mekky

5- Dorothy Guirgues Dawood

6- Mohammad Naser Ibrahim

Supervisor: Eng. Khaled El-Liethy

Agenda Overview

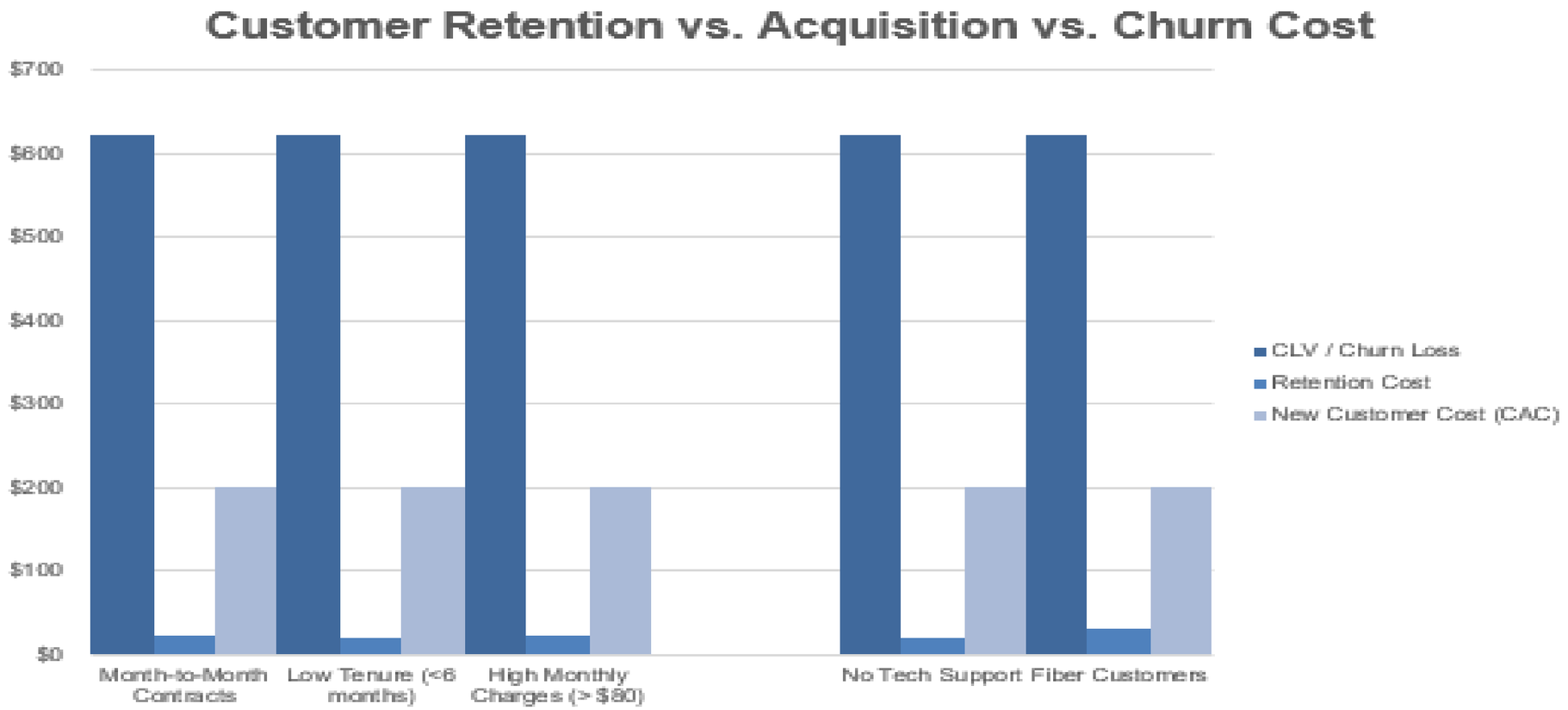
1. Introduction

Customer churn is a major challenge for telecom companies, as losing customers directly **reduces revenue** and increases **acquisition costs**. Predicting churn enables proactive **retention strategies**, helping maintain **loyalty** and **maximize revenue**. This project implements a full **machine learning pipeline**, from **data cleaning** to **deployment**, combining **technical rigor with business understanding** to identify **patterns behind churn** and provide **actionable insights**.

2.Driver Model

Driver	Avg Monthly Profit	CLV / Churn Loss	Retention Offer	Retention Cost	New Customer Cost (CAC)	Saving if Retained
Month-to-Month Contracts	$26	$624	15% discount * 2 months	$22	$200	$602
Low Tenure (<6 months)	$26	$624	Welcome Offer Free Tech Support for 2 months	$20	$200	$604
High Monthly Charges (> $80)	$26	$624	Bundle Offer $8 discount * 3 months	$24	$200	$600
No Tech Support	$26	$624	Free Tech Support $10 * 2 month	$20	$200	$604
Fiber Customers	$26	$624	$10 discount * 3 months	$30	$200	$594

2.Driver Model



To maximize profitability, the company should immediately direct its investment to high ROI (30 times) retention efforts, as they are significantly more cost-effective than acquiring new customers.

3. Project Pipeline

ANALYZE CUSTOMER BEHAVIOR

1

- Activity: Identify key churn drivers
- Output: Deep understanding of the problem

DATA PREPARATION AND FEATURE ENGINEERING

2

- Activity: Preprocess, engineer and select most relevant features
- Output: Clean and structures modeling dataset

TRAIN AND COMPARE CLASSIFICATION MODEL

3

- Activity: Train and compare six classification models
- Models: Logistic Regression, Random Forest, Gradient Boosting, XGBoost, LightGBM and CatBoost

MODELS PERFORMANCE EVALUATION

4

- Activity: Evaluate Models performance to identify the optimal predictive solution
- Evaluation Metrics: Accuracy, Precision, Recall, F1-score and AUC
- Output: The Optimal Model

DERIVE ACTIONABLE BUSINESS INSIGHTS

5

- Activity: Translate the winning model output into strategic recommendations
- Output: Effective Actionable Business Insights to guide retention strategies

4. Stakeholder Analysis

Stakeholder	Role	Needs/Expectations
Telecom Management	Decision Makers	Accurate churn predictions
Customer Service Team	Implement Retention Actions	Clear insights for high-risk customers
Data Science Team	Model Development	Clean data and reliable model
IT / DevOps	Deployment	Easy integration and scalable system
Customers	Indirectly affected	Improved service and offers

5. Dataset Overview

Dataset Name: Telco Customer Churn Dataset

Source: Kaggle – Telco Customer Churn

Dataset Design:

Column	Type	Description	Duplicates	Nulls
customerID	String	Unique identifier	0	0
gender	String	Customer gender	0	0
seniorCitizen	Int	Binary flag	0	0
partner	Int	Binary flag	0	0
dependents	Int	Binary flag	0	0
tenure	Int	Months as customer	0	0
contract	String	Contract type	0	0
paymentMethod	String	Payment method	0	0
monthlyCharges	Float	Monthly subscription	0	0
totalCharges	Float	Total charges	0	0
churn	Int	Target variable	0	0

Churn rate: ~26.5% (moderately imbalanced)

6. Data Preprocessing & Cleaning

0

Duplicates

TotalCharges

(from categorical to numeric)

Data Type Conversion

11

(appears after Data Type Conversion)

Missing Values Removed from TotalCharges

Outliers:

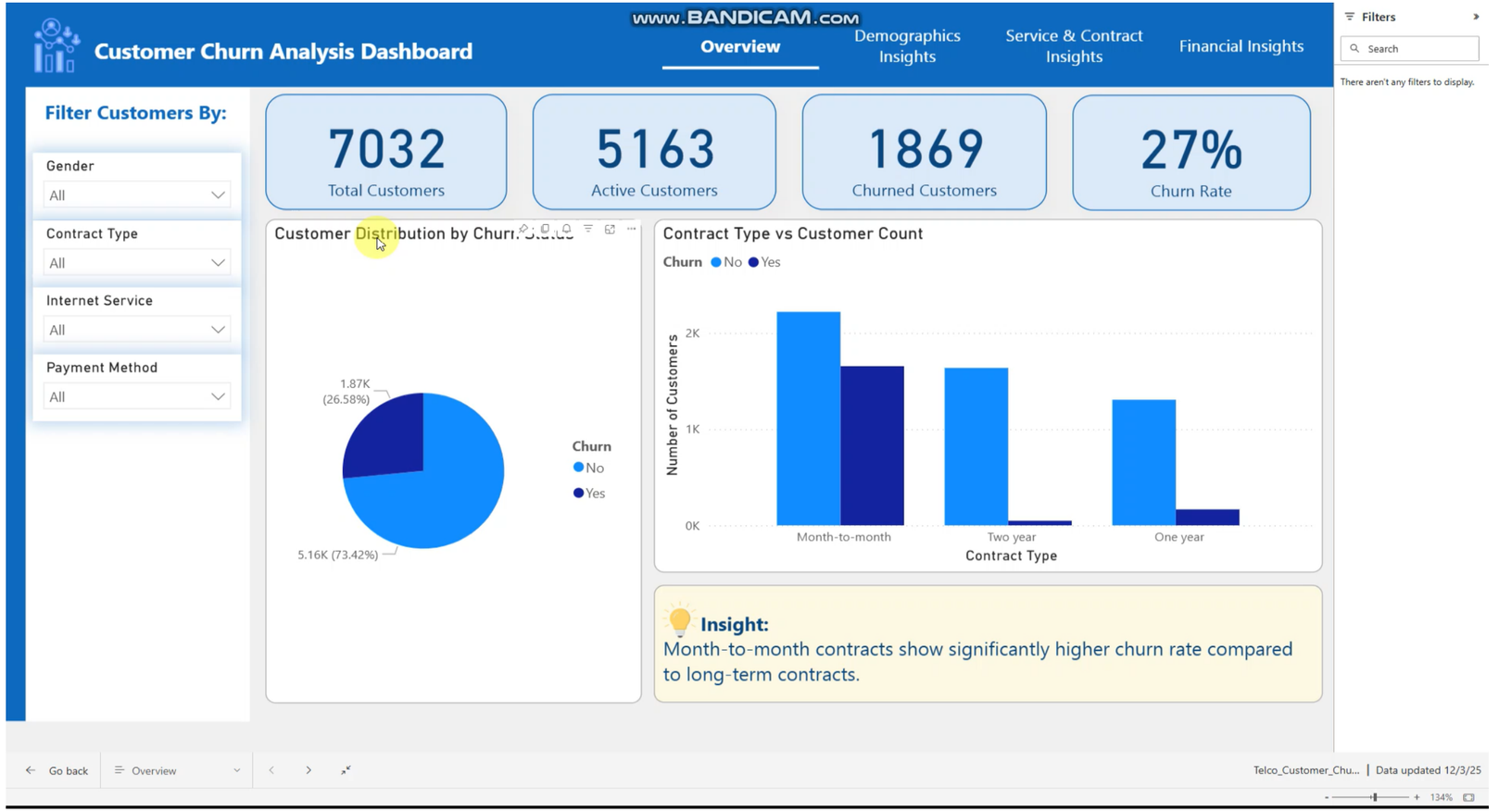
- Numeric features inspected
- No extreme values removed (represent real customer behavior)

Encoding: Applied only for models that cannot handle categorical features

Dataset: Saved for reproducibility: cleaned_Telco-Customer-Churn.csv

7. Exploratory Data Analysis (EDA)

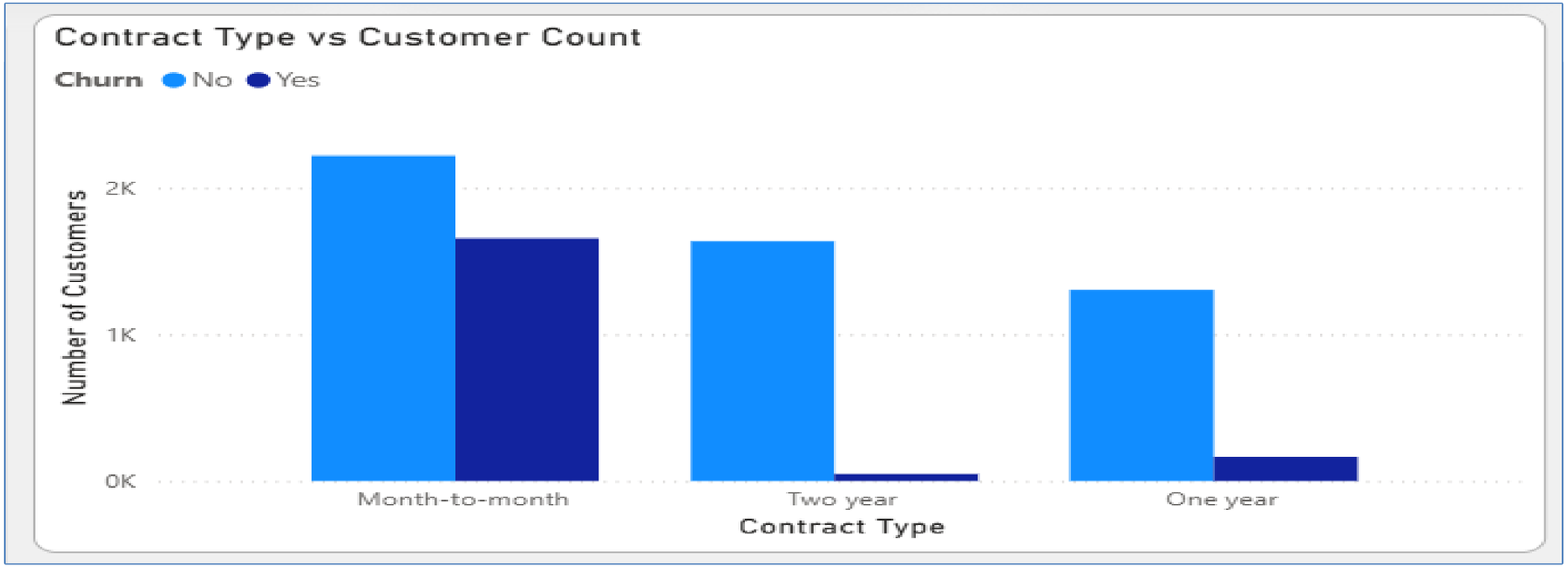
7.1: Dashboard Demo:



Note: We performed the exploratory analysis using both Python (Plotly) and Power BI.
For this presentation, Power BI visuals are used as they provide clearer and more intuitive business insights.

7. Exploratory Data Analysis (EDA)

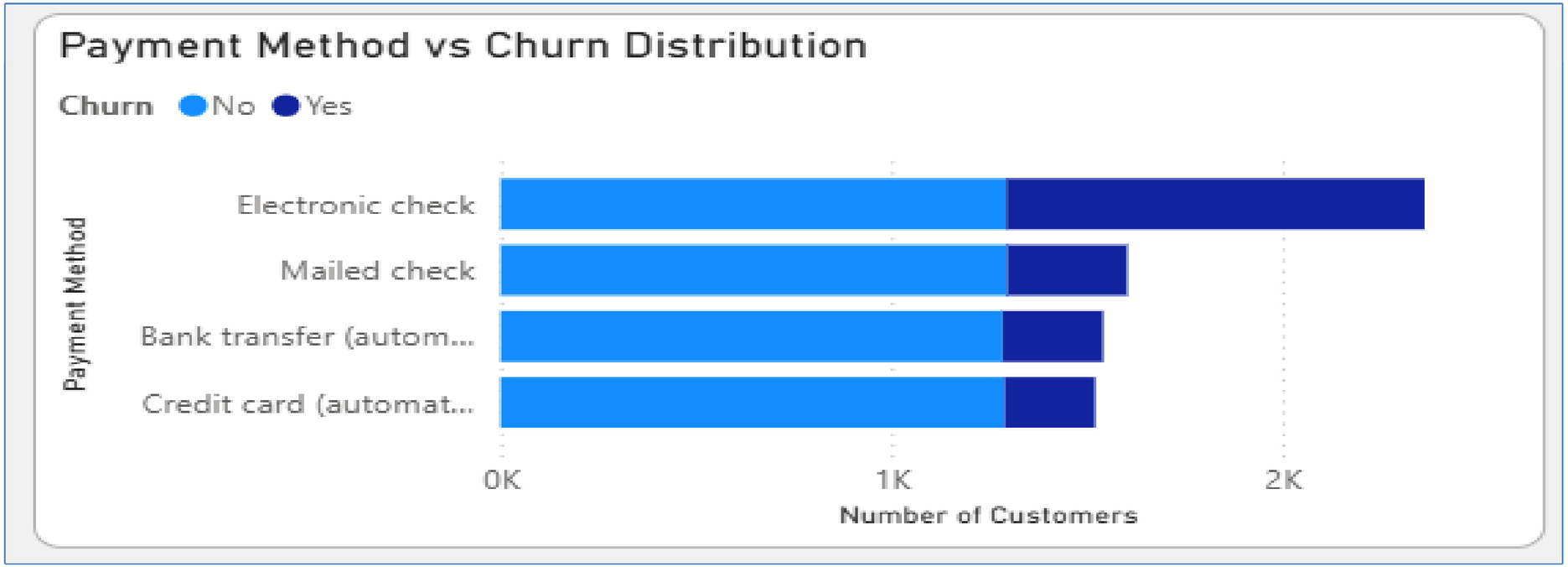
7.2.1 Main Insights:



Month-to-month customers have the highest churn rate, making contract type one of the strongest churn drivers.

7. Exploratory Data Analysis (EDA)

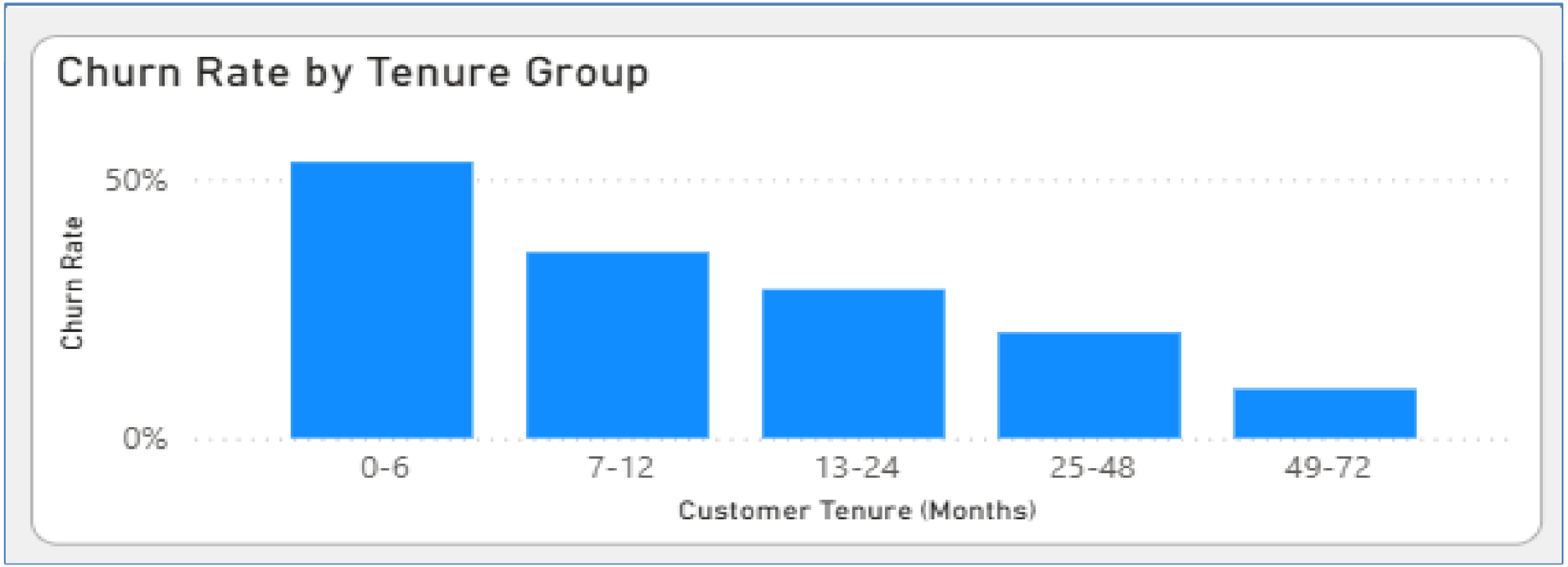
7.2.2 Main Insights:



Customers paying through Electronic Check churn significantly more than those using other payment methods.

7. Exploratory Data Analysis (EDA)

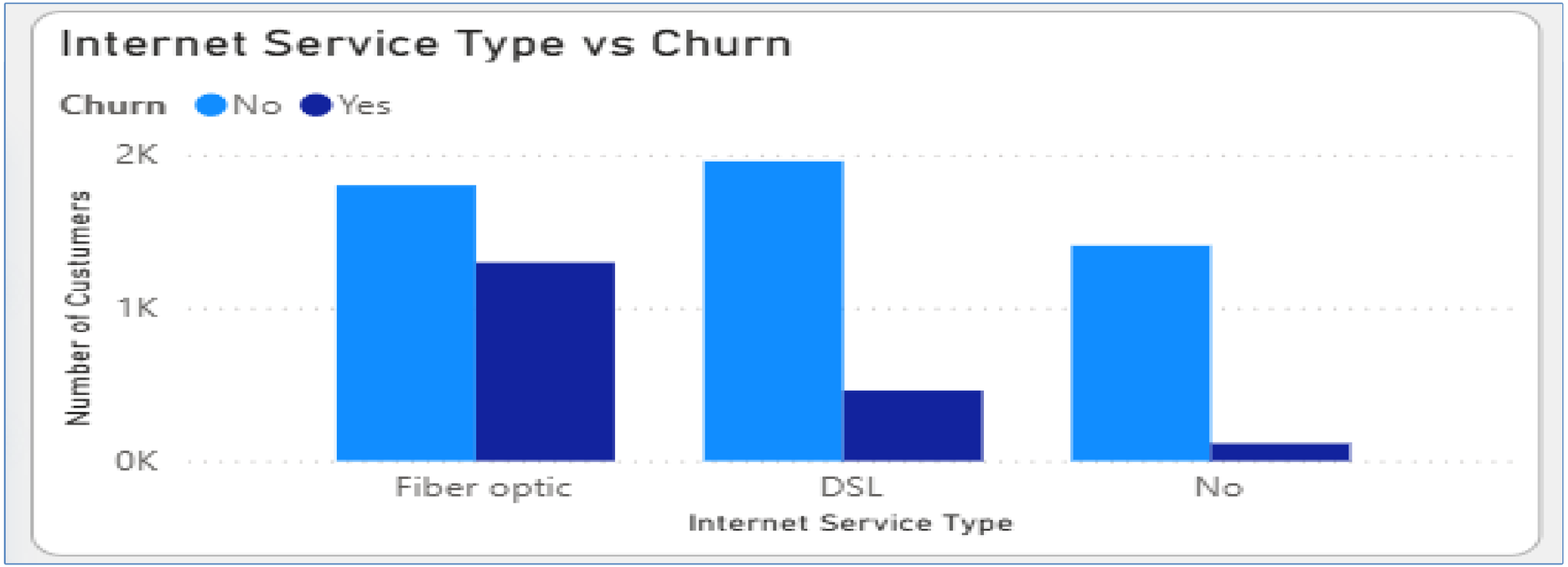
7.2.3 Main Insights:



Short-tenure customers (0–12 months) contribute the largest portion of churn, indicating early dissatisfaction.

7. Exploratory Data Analysis (EDA)

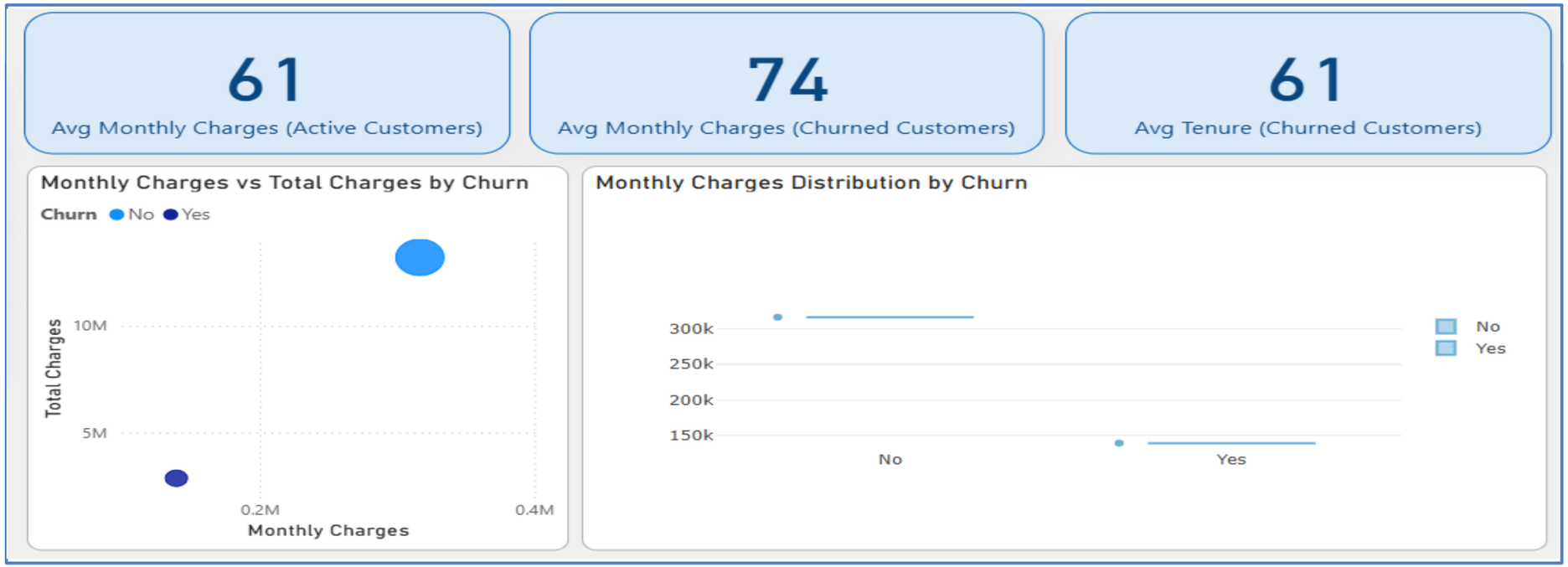
7.2.4 Main Insights:



Fiber Optic users show higher churn rates than DSL customers, suggesting potential service or pricing issues.

7. Exploratory Data Analysis (EDA)

7.2.5 Main Insights:



Customers with higher monthly charges are more likely to churn, while those with lower monthly charges and longer tenure show stronger loyalty

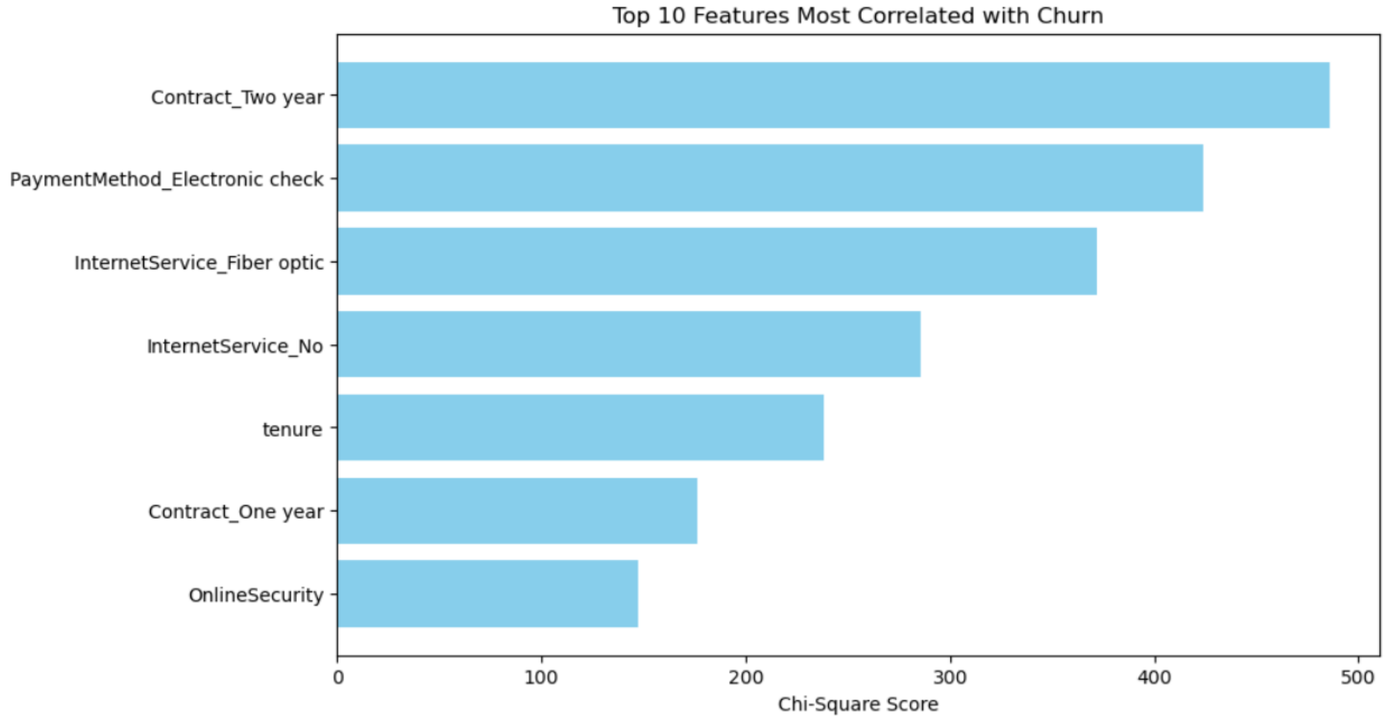
8. Statistical Tests & Feature Selection(1)

Chi-Square Test (Categorical Features)
Significant association with churn ($p < 0.05$):

- **Contract**
- **PaymentMethod**
- **InternetService**

Mann–Whitney U Test (Numeric Features)
Significant differences between churn vs non-churn groups:

- **Tenure**
- **MonthlyCharges**



8. Statistical Tests & Feature Selection (2)

1- Selected categorical columns.
2- Created a contingency table using pd.crosstab().
3- Applied chi2_contingency() from SciPy.
4- Interpreted the p-value (p < 0.05 → significant association).

Significant features found:

Contract, Payment Method, Internet Service

Contract Type	Churn = Yes	Churn = No
Month-to-month	1654	2220
One-year	130	850
Two-year	70	1650

Internet Service	Churn = Yes	Churn = No
Fiber optic	1100	1500
DSL	300	1400
No	71	850

Payment Method	Churn = Yes	Churn = No
Electronic Check	1071	1300
Mailed Check	200	1394
Bank Transfer	300	1290
Credit Card	270	1525

9. Feature Engineering & Encoding(1)

9.1 Feature Engineering:

1. Tenure Grouping – Data-Driven Binning Approach

- Did not use arbitrary or equal-width binning.
- Analyzed the distribution of tenure and its relationship with churn.
- Churn rates were evaluated across multiple tenure intervals (6 months, 12 months, quartiles, and natural breaks)
- Clear behavioral changes in churn were observed at specific tenure points
- Based on these patterns, tenure was grouped into lifecycle-based bins:
- 0–12 months (early-stage customers – highest churn) - ~32%
- 13–24 months (medium-term customers) - ~16%
- 25–48 months (established customers) - ~9%
- 49–72 months (long-term loyal customers) - ~4%
- This data-driven binning captures real customer behavior and improves model interpretability and performance.

2. AvgChargesPerMonth

- New feature calculated as:
  AvgChargesPerMonth = TotalCharges / Tenure
- Represents the true average monthly spending for each customer.

3. ServiceCount

- Counted the number of active services per customer.
- Useful for analyzing relationships between service bundles and churn.

4. Interaction Feature

- Created an interaction term:
  Tenure × MonthlyCharges
- Highlights customers with low tenure but high monthly charges, who are typically more likely to churn.

9. Feature Engineering & Encoding(2)

9.2 Encoding:

1. Binary Encoding

- Identified binary categorical features (Yes / No)
- Converted values to numeric format
  Yes → 1, No → 0
- Ensured features are stored as integers for model compatibility.
- Improved model interpretability and training efficiency

2. One-Hot Encoding

- Selected multi-category categorical features
  Contract, PaymentMethod, InternetService
- Applied One-Hot Encoding using pd.get_dummies()
- Created separate binary columns for each category
- Used drop_first=True to avoid multicollinearity
- Prevented introducing unintended ordinal relationships

"*Binary Encoding* is suitable when a feature has *only two possible values*, while *One-Hot Encoding* is necessary for features with *multiple categories* to avoid incorrect data representation."

10. Modeling & Evaluation (1)

Model	Accuracy	AUC	Precision (Class 0)	Recall (Class0)	F1 (Class0)	Precision (Class 1)	Recall (Class1)	F1 (Class1)
Logistic Regression	0.72	0.82	0.90	0.71	0.79	0.49	0.77	0.60
Random Forest	0.77	0.80	0.82	0.90	0.85	0.60	0.44	0.51
Gradient Boosting	0.72	0.82	0.90	0.71	0.79	0.49	0.77	0.60
XGBoost	0.77	0.81	0.83	0.88	0.85	0.60	0.50	0.55
LightGBM	0.78	0.82	0.84	0.87	0.85	0.60	0.55	0.57
CatBoost	0.78	0.84	0.89	0.81	0.85	0.57	0.72	0.64

CatBoost was chosen for final deployment

AUC & Recall were Chosen as the Primary Evaluation Metrics

10. Modeling & Evaluation (2)

Why CatBoost was chosen for final deployment?

- **Native handling of categorical features** (no one-hot encoding needed)
- **High performance** on tabular customer data
- **Reduced overfitting** using ordered boosting
- **Effective handling of imbalanced data** using class weights
- **Stable probability predictions** suitable for churn ranking and threshold tuning

10. Modeling & Evaluation (3)

Why AUC & Recall Were Chosen as the Primary Evaluation Metrics?

1- Why AUC matters?

- Measures separation between churners and non-churners
- Threshold-independent
- Evaluates probability quality

2- Why Recall matters?

- Recall = % of churners we successfully detect
- Minimizes “Missed churners” → highest business cost
- Ensures the model identifies at-risk customers

10. Modeling & Evaluation (4)

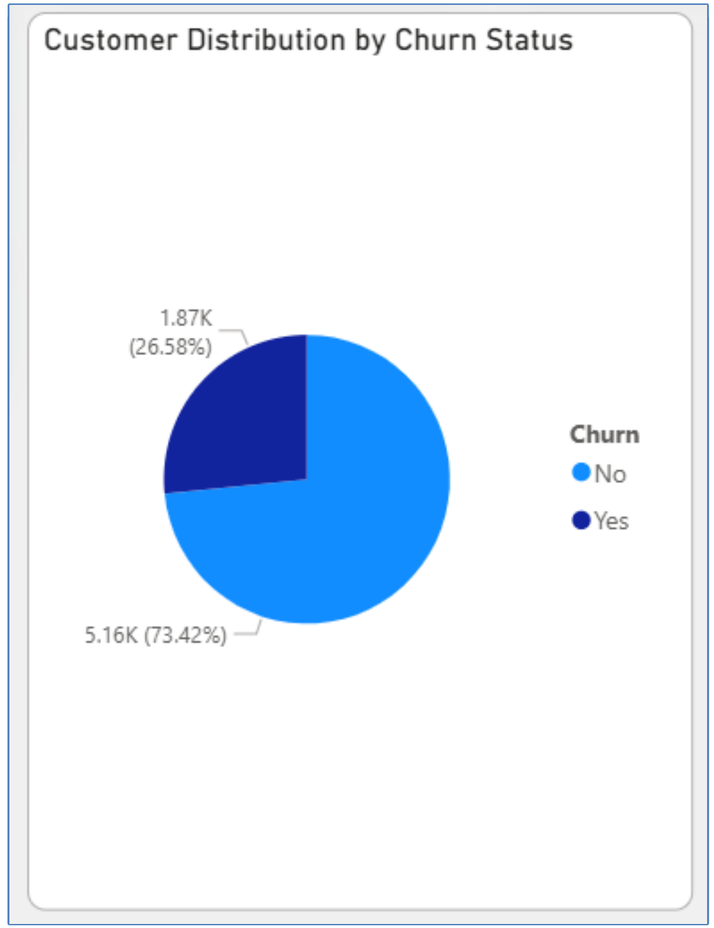
Handling Class Imbalance:

Dataset Issue:

Class	Count	Description
0 – Not Churned	High	Majority class
1 – Churned	Low	Underrepresented, harder to predict

Our Solution:

- Used class_weights instead of SMOTE
- Avoided generating synthetic customer records
- Made the model penalize misclassifying churners more heavily
- Increased recall & maintained stable AUC



10. Modeling & Evaluation (5)

Model Training Setup:

CatBoost Configuration:

Parameter	Value
Iterations	600
Depth	5
Learning Rate	0.07
Loss Function	Logloss
Eval Metric	AUC
Early Stopping	80 rounds
Bootstrap Type	Bayesian
Bagging Temperature	2
Class Weights	Applied

Why this Setup?

- Balanced generalization and model stability
- Reduced overfitting through Bayesian bootstrapping
- Faster and more consistent convergence
- Strong separation power measured by AUC
- High sensitivity to churners through optimized Recall

10. Modeling & Evaluation (6)

Thresholds Tested:

Threshold	Accuracy	Precision	Recall	F1
0.45	0.711	0.475	**0.837**	0.606
0.50	0.738	0.505	0.810	0.622
0.55	0.763	0.539	0.754	0.629
0.60	**0.783**	**0.574**	0.719	**0.638**

Why this matters?

💡 Lower threshold → **Higher Recall** (catch more churn)

💡 Higher threshold → **More precision** & **higher accuracy**

💡 We selected **0.60** because it gives the best **balance** between **AUC, Recall, and F1**

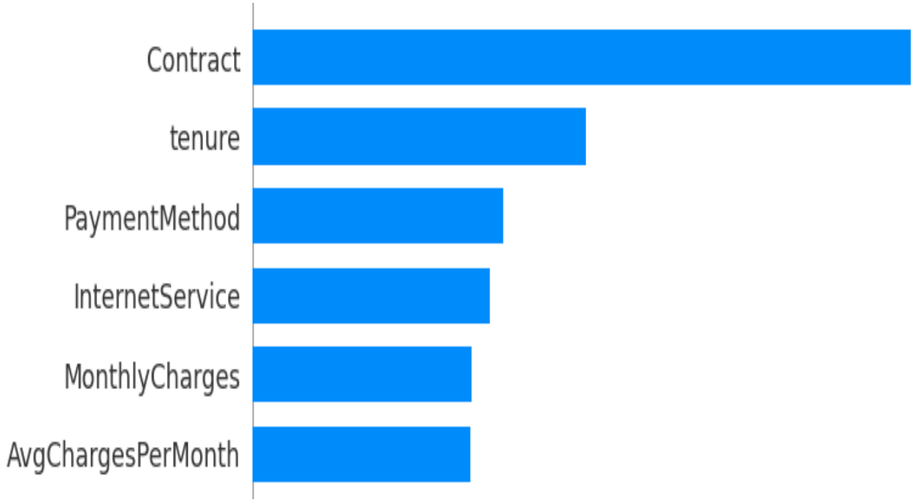
10. Modeling & Evaluation (7)

SHAP Feature Importance:

What SHAP reveals?

- Contribution of each feature to churn probability
- Transparent explanation the business can trust
- Helps prioritize retention actions

Top Drivers of Churn



11. Business Impact & Recommendations

Early churn prediction enables targeted retention campaigns, allowing the company to proactively address high-risk customers.

Using CatBoost along with recommended actions can reduce churn by ~6%, improving revenue stability and customer lifetime value.

Key churn drivers identified:

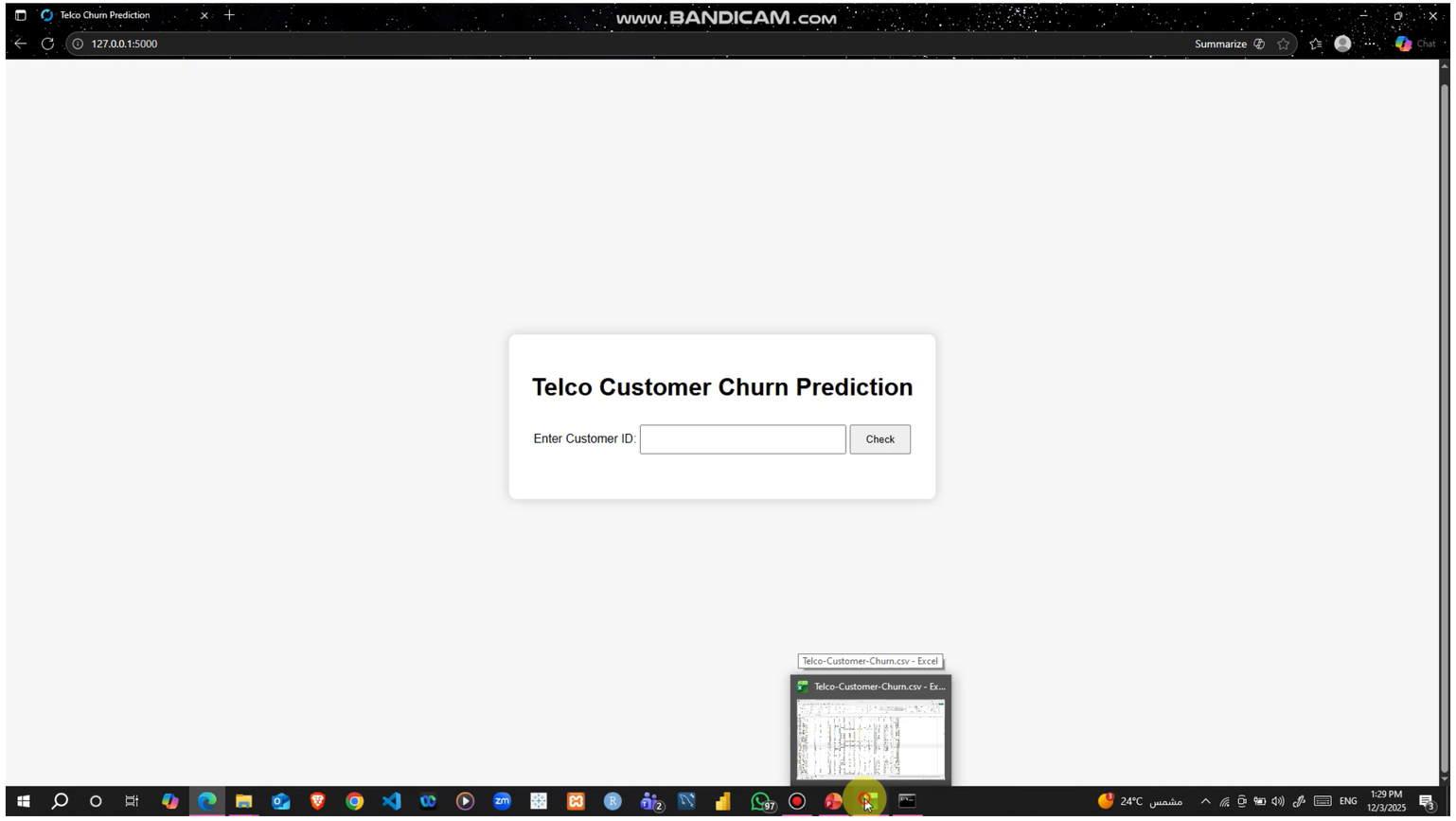
- Contract Type
- Payment Method
- Customer Tenure
- Service Usage Patterns

Recommended Actions:

- Offer personalized retention incentives for month-to-month or high-risk customers.
- Promote automatic payment methods to reduce churn.
- Focus support and engagement on short-tenure and high-spending customers.
- Enhance services for Fiber Optic users and senior citizens, addressing specific needs.

12. Web Application(1)

Demo:



- Enter Customer ID → show churn prediction & probability
- Invalid Customer IDs are detected and appropriate error messages are returned without crashing the system

12. Web Application(2)

Implementation / Tools:

Backend

Frontend

ML Model

13. Conclusion

In this project, we built a complete **machine learning pipeline** to predict **customer churn** for a telecom company using **Python**, **Power BI**, and **CatBoost**. We identified the **main factors driving churn**, engineered **meaningful features**, and created a **web application ready for deployment**. The model achieved strong **predictive performance** (**AUC ≈ 0.84**, **high recall for churners**) and provides **interpretable insights** through **SHAP feature importance**. By predicting churn **early** and applying **targeted retention strategies**, the company can **reduce churn by around 6%**, **stabilize revenue**, and **increase customer lifetime value**. This solution combines **technical accuracy** with **practical business insights**, making it both **actionable** and **reliable**.

Thank You