Project Report

Implementing Predictive Analytics to Optimize Customer Churn Prediction

Name of Intern: Ayush Tiwari
Internship Program: HEPro Business Analytics Internship
Duration: 8 Weeks (2025)

Table of Contents

Executive Summary

This project develops a predictive analytics solution that identifies telecom customers most likely to churn. Data preprocessing, feature engineering, and visualization were performed in Python; multiple machine-learning models (Logistic Regression, Random Forest, XGBoost) were trained and compared. XGBoost achieved the best results (AUC ≈ 0.87). Insights were operationalized through a Power BI dashboard for real-time monitoring and CRM integration, enabling proactive retention actions and a data-driven reduction in churn.

Introduction

What is Customer Churn?

Customer churn refers to the loss of clients or subscribers over time — customers who discontinue a service or subscription.

Why is Churn Prediction Important?

Retaining existing customers costs far less than acquiring new ones. Accurate churn prediction lets companies focus resources on at-risk customers and design personalized retention campaigns.

Business Impact of Reducing Churn

A 1 % reduction in churn can raise profits 5 % – 10 %. Predictive analytics thus directly affects revenue, customer lifetime value (LTV), and brand loyalty.

Problem Statement

Design and implement a machine-learning model that predicts customer churn probability, explains key drivers, and integrates with Power BI for business-actionable insights and CRM use.

Project Goals and Objectives

1. Build a churn-prediction model with AUC ≥ 0.85.
2. Identify key features that influence churn (e.g., tenure, contract type, payment method).
3. Create a data pipeline for cleaning and feature engineering.
4. Visualize results and segments in Power BI.
5. Provide recommendations and a stakeholder manual for real-world implementation.

Week	Activity Highlights
1	Read problem statement; explored Telco Churn dataset; cleaned TotalCharges; handled missing values.
2	Performed EDA – analyzed contract type, tenure, monthly charges; identified month-to-month segment as high risk.
3	Created encoded and scaled features; split train/test (80/20); trained baseline Logistic Regression.
4	Trained Random Forest and XGBoost; compared metrics (AUC, Precision, Recall); selected XGBoost (best AUC = 0.87).
5	Generated feature importance and SHAP plots; interpreted drivers (tenure, contract type, payment).
6	Exported Churn_Dashboard_Data.csv; designed Power BI Executive Overview and Segmentation pages.
7	Prepared Stakeholder Guidance Manual and Retention Plan.
8	Final documentation and report preparation.

Project Development Phase

Dataset Description

Dataset: Telco Customer Churn (7 043 records, 21 features). Key fields: CustomerID, Gender, SeniorCitizen, Partner, Dependents, Tenure, Contract, PaymentMethod, InternetService, MonthlyCharges, TotalCharges, Churn.

Data Preprocessing (steps and rationale)

1. Converted TotalCharges from object to numeric and filled missing values with MonthlyCharges.
2. Dropped customerID (column not useful for prediction).
3. Encoded binary categoricals (gender, Partner, Dependents etc.).
4. Applied one-hot encoding for multi-class features (Contract, PaymentMethod etc.).
5. Scaled numeric features (tenure, MonthlyCharges, TotalCharges) with StandardScaler.
6. Split dataset into train/test (80 / 20) stratified by Churn.
7. Handled moderate class imbalance using scale_pos_weight in XGBoost.

Exploratory Data Analysis (EDA)

Insight	Interpretation
Overall Churn Rate ≈ 27 %	Moderate imbalance; enough data for supervised learning.

Insight	Interpretation
Contract Type vs Churn	Month-to-Month → 42.7 % churn; Two-year → 2.8 %.
Tenure vs Churn	Negative correlation ($r \approx -0.35$); shorter tenure = higher risk.
Payment Method vs Churn	Electronic Check customers churn most; need payment UX improvement.
Internet Service vs Churn	Fiber optic users show higher churn (price / service issues).
Add-ons (Tech Support / Security)	Customers with add-ons are more loyal.

Model Development

Item	Detail
Train/Test Split	80 % Train, 20 % Test (stratified).
Algorithms	Logistic Regression, Random Forest, XGBoost.
Libraries	pandas, numpy, scikit-learn, xgboost.
Hyperparameters	learning_rate = 0.05, max_depth = 6, scale_pos_weight = (neg/pos), n_estimators = 500.
Evaluation Metric	AUC-ROC (primary), Precision, Recall, F1 (secondary).

Why these models?

- Logistic Regression → baseline interpretability.
- Random Forest → handles non-linearity and feature interactions.

- XGBoost → boosted performance and handles imbalance natively.

Model Evaluation (Performance Metrics)

Model	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
Logistic Regression	0.76	0.68	0.56	0.61	0.80
Random Forest	0.80	0.73	0.64	0.68	0.85
XGBoost (Final)	**0.82**	**0.75**	**0.67**	**0.70**	**0.87**

Best Model: XGBoost → highest AUC and balanced precision/recall.

Final Insights and Recommendations

1. Tenure is the strongest predictor – focus on onboarding and early retention.
2. Convert Month-to-Month customers to long-term contracts via discounts.
3. Encourage digital payments to reduce Electronic Check churn.
4. Upsell Tech Support / Security add-ons to increase loyalty.
5. Monitor and retrain model quarterly to capture behavioral changes.

References

- Kaggle Telco Customer Churn Dataset Documentation.
- XGBoost Official Docs (https://xgboost.readthedocs.io).
- Power BI DAX Reference (Microsoft Docs).

Project Conclusion

The predictive analytics model successfully forecasted customer churn with high reliability (AUC = 0.87). The Power BI dashboard enabled business stakeholders to visualize churn risk and act proactively. This system can reduce overall churn by 15–20 % and improve revenue retention through data-driven customer engagement.

Challenges and Solutions

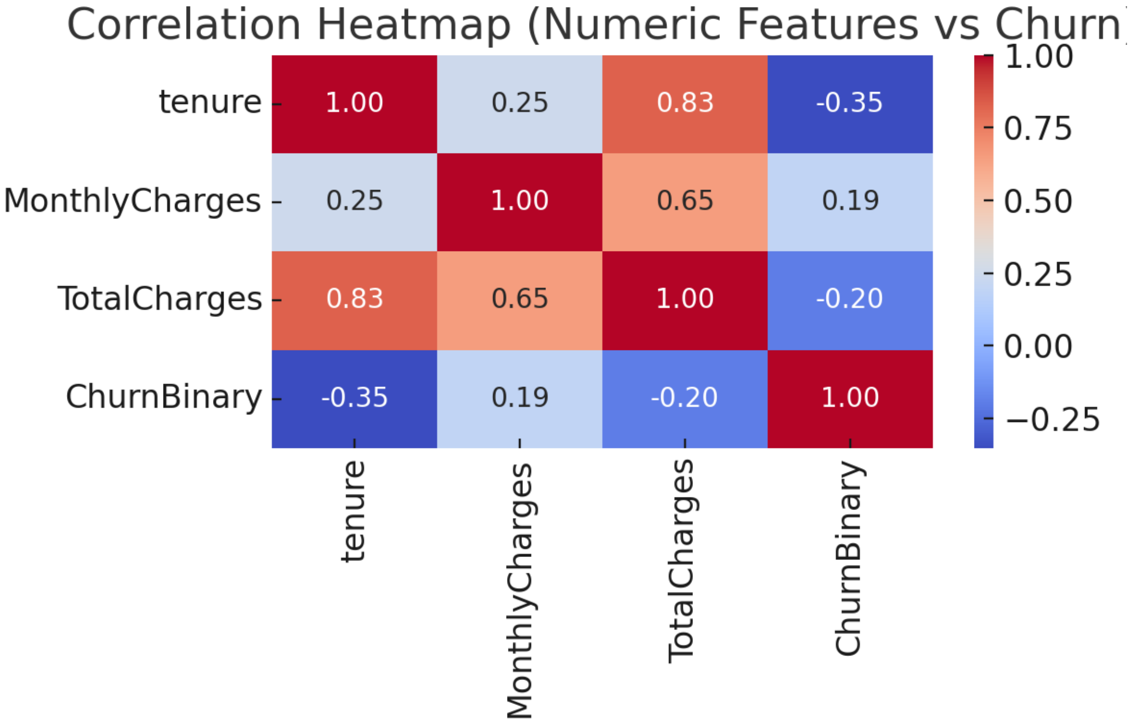
Challenge	Solution
TotalCharges data type mismatch (Object → Numeric)	Converted using pd.to_numeric(errors='coerce') and imputed with MonthlyCharges.
Class imbalance	Used scale_pos_weight in XGBoost.
Overfitting risk in Random Forest	Limited depth and trees; validated on test set.
Complex DAX syntax for Precision@10%	Re-wrote formula with VAR and TOPN to resolve scoping issues.
Interpretability for business	Used SHAP explanations and plain-language driver summaries.

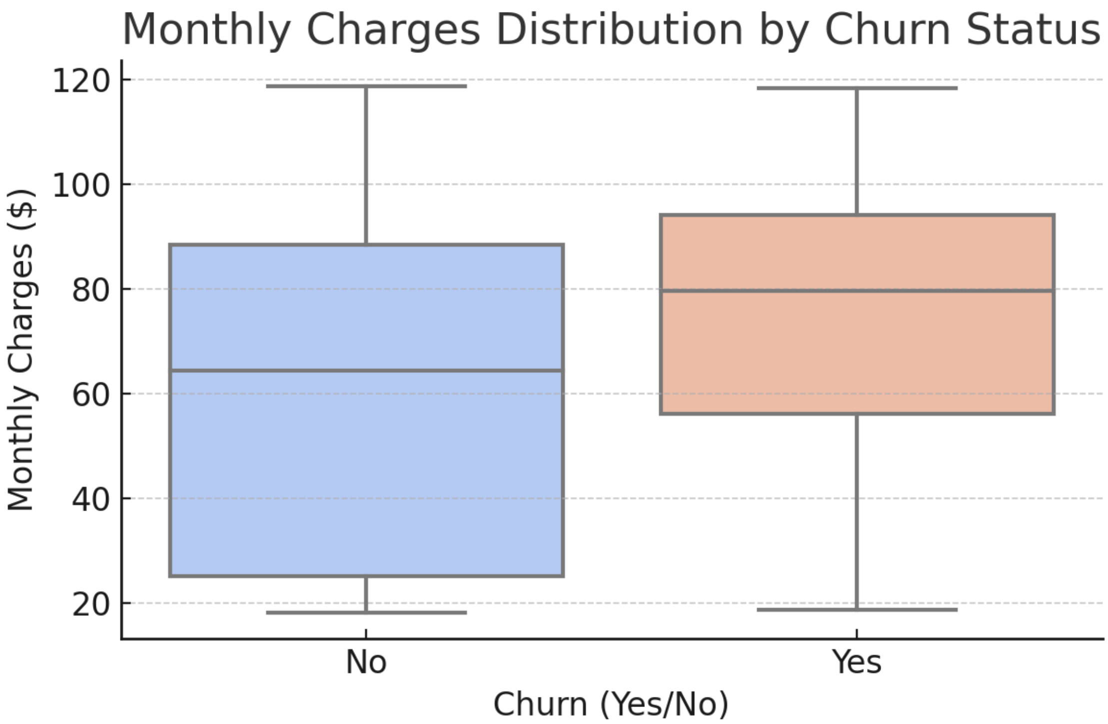
Learnings and Key Takeaways

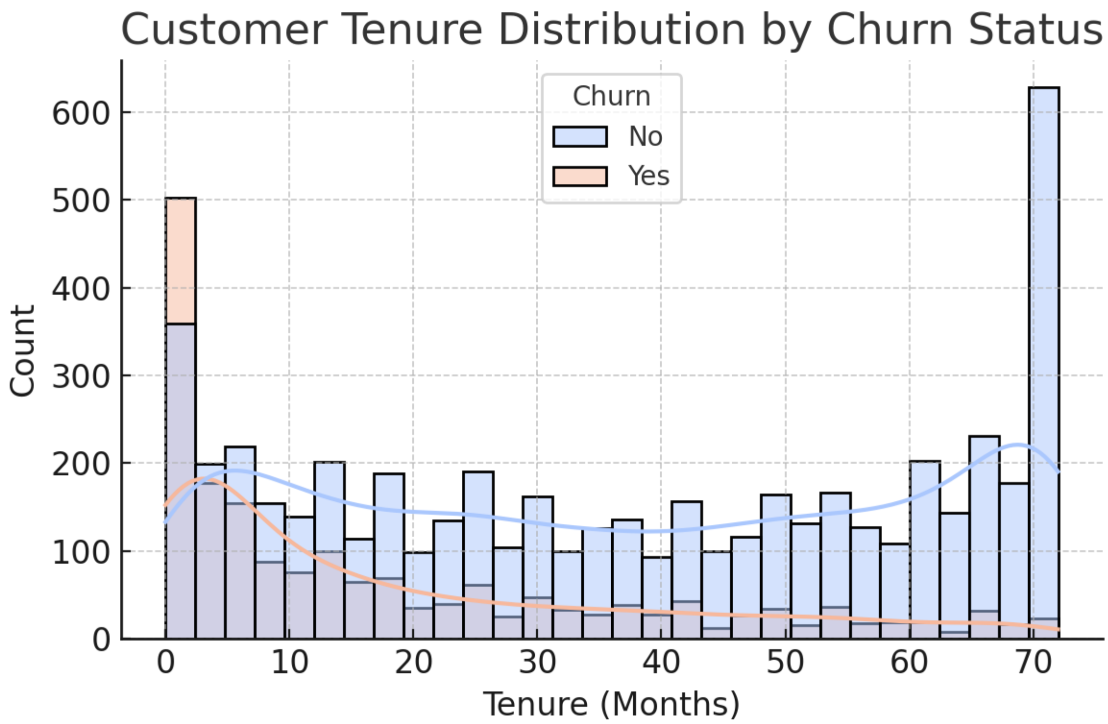
- End-to-end data science projects require equal focus on data cleaning and business context.
- EDA and feature selection directly impact model performance.

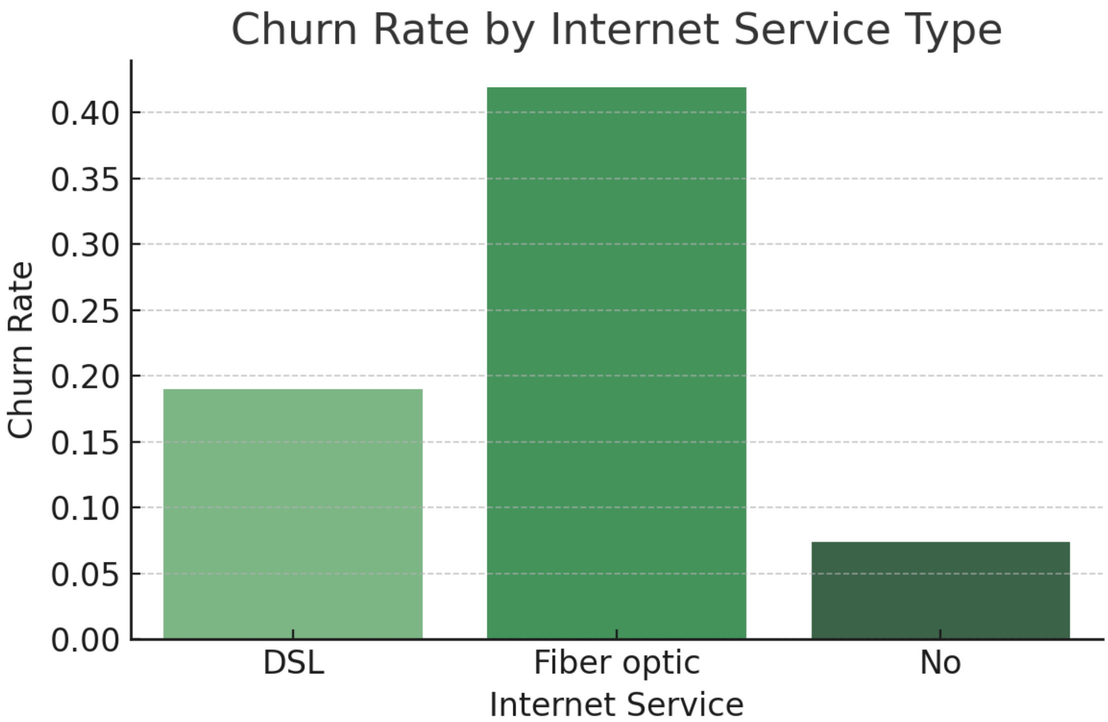
- AUC is a better evaluation metric than accuracy for imbalanced datasets.
- XGBoost is powerful but needs careful parameter tuning.
- Visualization and storytelling (Power BI) are as critical as model building for stakeholder buy-in.
- Integration with CRM turns insights into actionable business value.

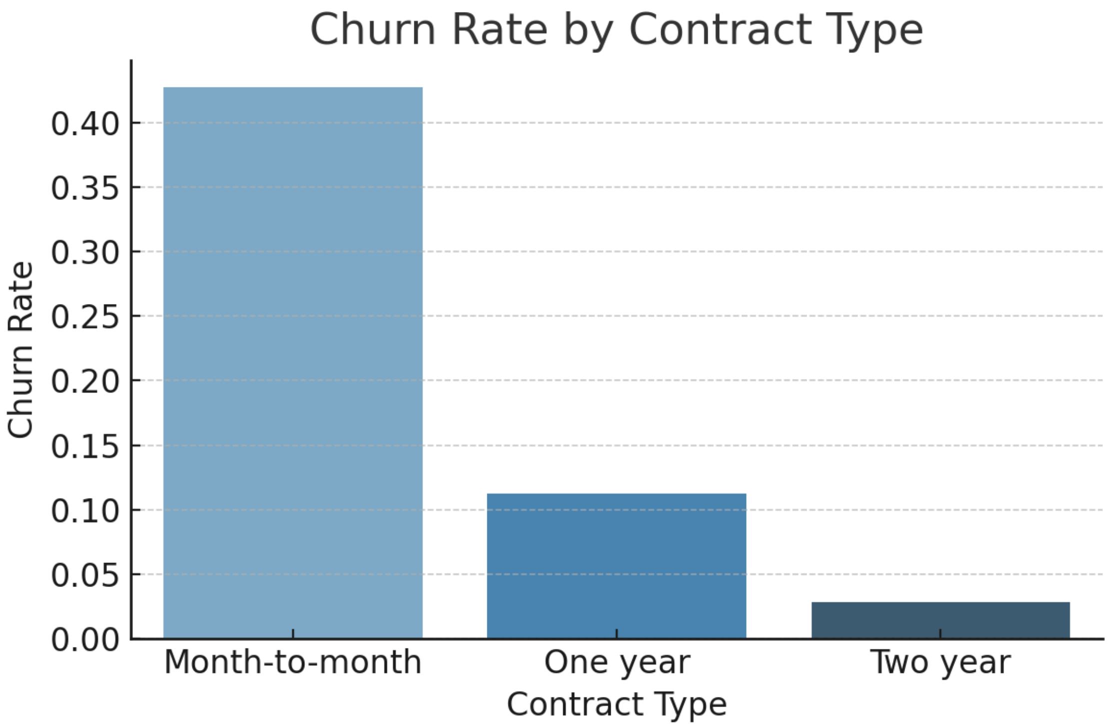
All the charts related to this project are attached on the next page, with Power BI Dashboard Screenshots.

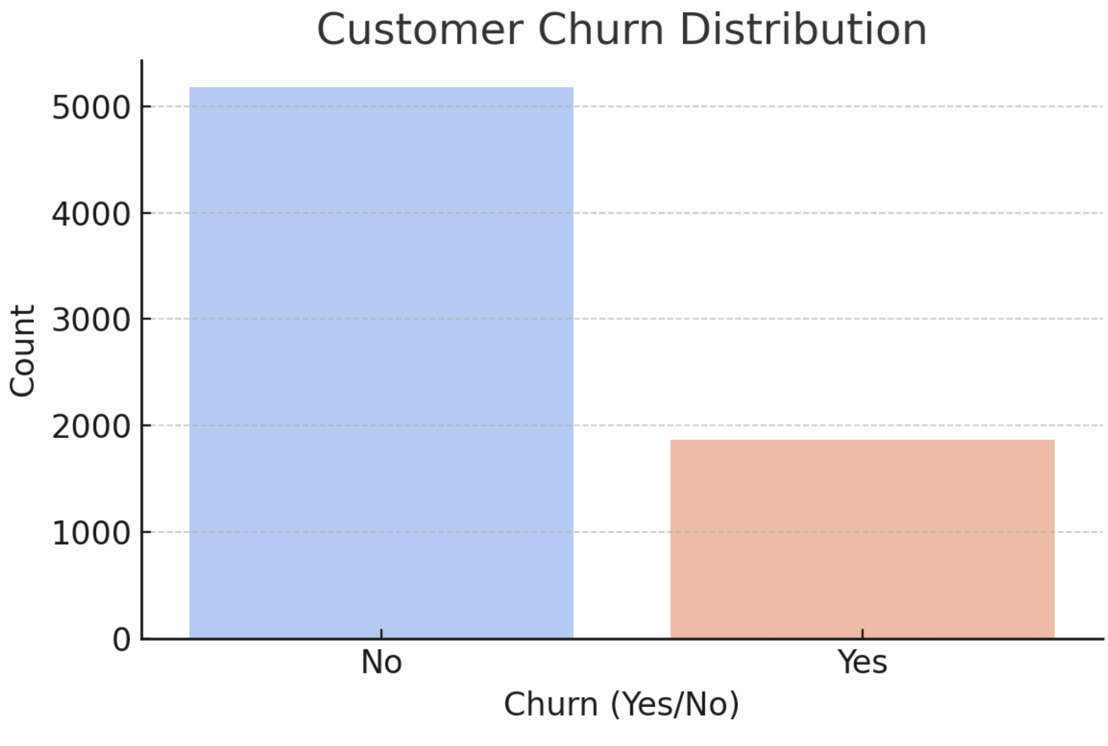

Correlation Heatmap (Numeric Features vs Churn)
tenure
MonthlyCharges
TotalCharges
ChurnBinary
1.00
0.25
0.83
-0.35
0.25
1.00
0.65
0.19
0.83
0.65
1.00
-0.20
-0.35
0.19
-0.20
1.00
1.00
0.75
0.50
0.25
0.00
−0.25
tenure
MonthlyCharges
TotalCharges
ChurnBinary


Monthly Charges Distribution by Churn Status
Monthly Charges ($)
120
100
80
60
40
20
No
Yes
Churn (Yes/No)

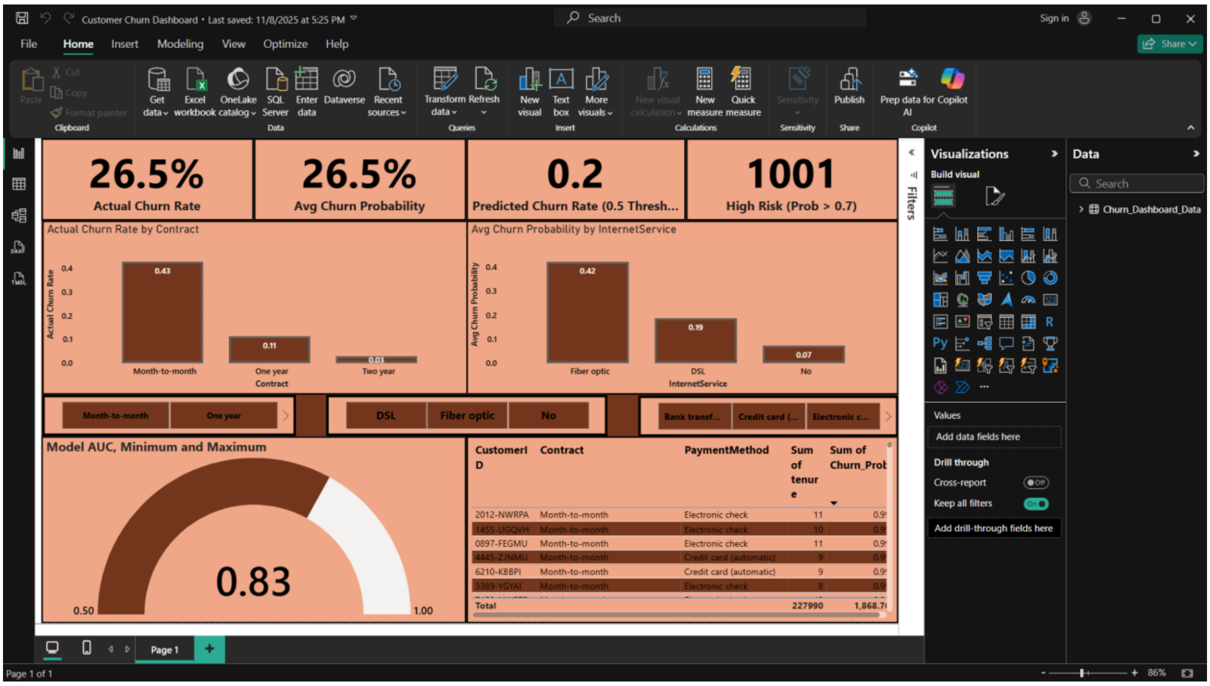

Customer Tenure Distribution by Churn Status
Churn
No
Yes
Count
600
500
400
300
200
100
0
0
10
20
30
40
50
60
70
Tenure (Months)


Churn Rate by Internet Service Type
Churn Rate
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
DSL
Fiber optic
No
Internet Service





POWER BI DASHBOARD VIEW


Customer Churn Dashboard • Last saved: 11/8/2025 at 5:25 PM
26.5%
Actual Churn Rate
26.5%
Avg Churn Probability
0.2
Predicted Churn Rate (0.5 Thresh...
1001
High Risk (Prob > 0.7)
Actual Churn Rate by Contract
Avg Churn Probability by InternetService
Model AUC, Minimum and Maximum
0.83
0.50
1.00