ChurnSight: Predicting and Preventing Customer Churn

Author: Kamran Habib
Tools: Azure Databricks (PySpark + MLlib) | Power BI (DAX + Visual Analytics)
Focus Areas: Data Engineering · Machine Learning · Data Visualization · Customer Analytics

---

1. Business Objective

Telecom providers lose significant revenue to customer churn.
ChurnSight was designed to detect churn risk, identify the main reasons behind it, and highlight high-value customers worth retaining using predictive analytics and interactive dashboards.

---

2. Technical Workflow

Stage	Platform	Key Tasks
Data Engineering	Azure Databricks	Ingested, cleaned, and transformed IBM Telco Churn dataset (7 K → 2 K valid rows). Handled missing TotalCharges, encoded categorical variables, built feature vectors.
Modeling	PySpark MLlib	Compared 4 algorithms — Logistic Regression, Decision Tree, Random Forest, GBT. Logistic Regression selected (AUC 0.858, F1 0.809).
Feature Enrichment	Databricks + PySpark	Added RFM (Recency-Frequency-Monetary) scores and Retention Priority metric = RFM × Churn Probability.
Visualization	Power BI	Connected final_scored.csv, created calculated columns and DAX measures for KPIs, designed interactive dashboards.

---

3. Power BI Dashboards

a. Data Overview

- KPIs: Total Customers = 2 004 | Churn Rate = 26.6 % | Avg Churn Probability = 26.7 % | Avg RFM = 8.9
- Purpose: Validate dataset and monitor overall risk exposure.

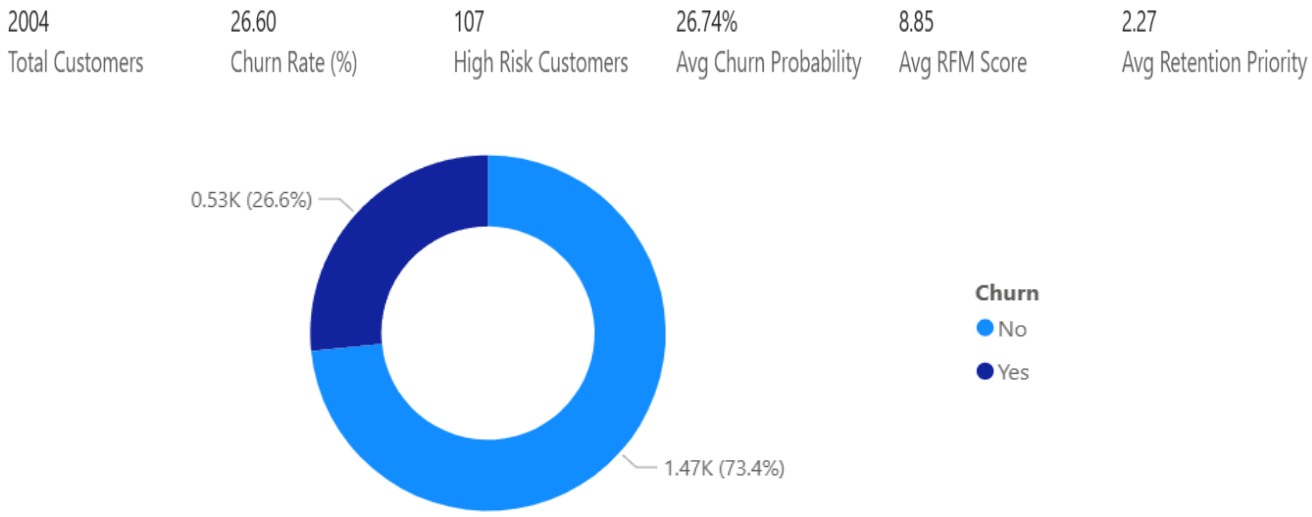

Figure 1 – Data Overview Cards & Customer Churn

---

b. Churn Analysis

- Month-to-Month contracts show the highest churn (~45 %).
- Fiber Internet and Electronic Check users are most volatile.
- Tenure > 24 months significantly reduces risk.

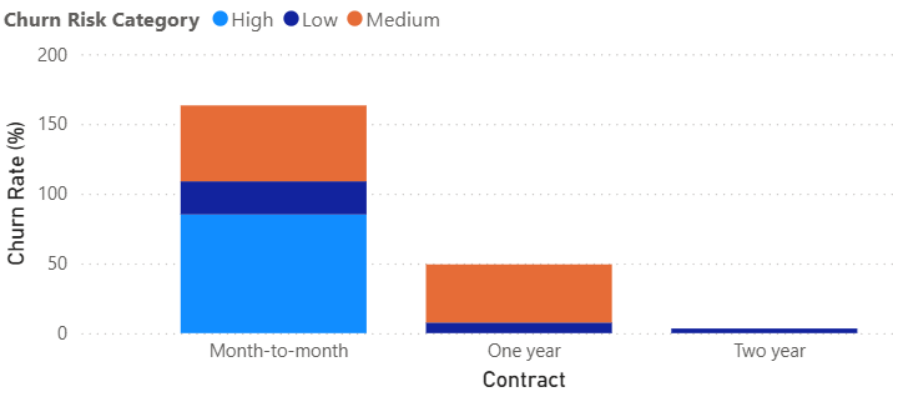

Figure 2. Contract vs Churn

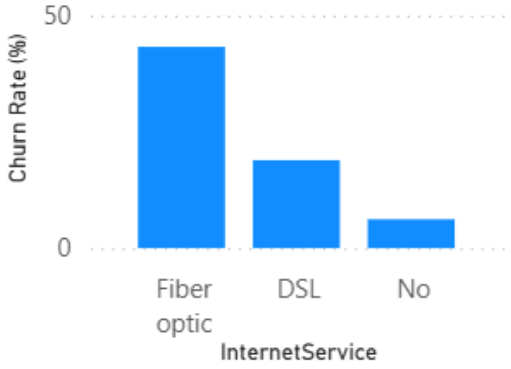

Figure 3. Internet Service vs Churn

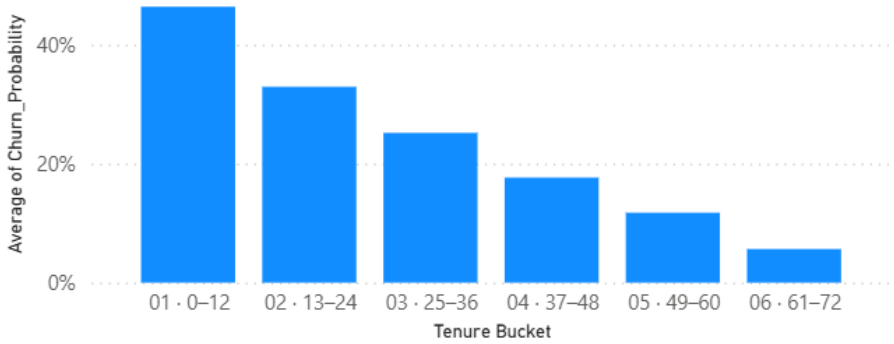

Figure 4. Tenure Bucket Trend

---

c. Retention & RFM Analysis

- Combined model probabilities with customer value scores.
- 107 customers flagged as high-risk & high-value.
- Top 20 customers listed for sales outreach.

RFM_Score (bins)	High	Low	Medium	**Total**
2		76	16	**92**
4	14	130	104	**248**
6	21	349	88	**458**
8	42	140	104	**286**
10	27	252	91	**370**
12	3	332	105	**440**
14		106	4	**110**
Total	**107**	**1385**	**512**	**2004**

Figure 5. RFM Matrix

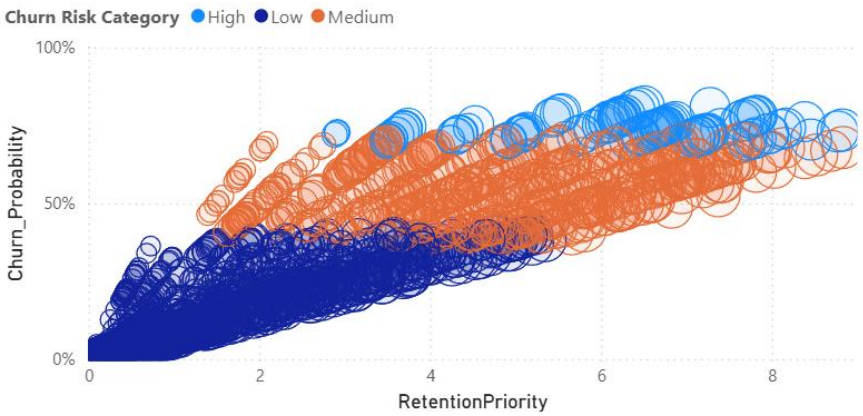

Figure 6. Retention Priority Scatter Plot

customerID	Contract	RFM_Score	Churn Probability %	RetentionPriority	MonthlyCharges
3606-TWKGI	Month-to-month	12	73.50	8.82	$106.9
3957-LXOLK	Month-to-month	13	67.90	8.83	$106.15
8868-WOZGU	Month-to-month	13	65.60	8.52	$105.7
7392-YYPYJ	Month-to-month	12	70.50	8.46	$100.65
3393-FMZPV	Month-to-month	13	66.90	8.69	$100.25
4817-VYYWS	Month-to-month	13	64.50	8.39	$100.2
9094-AZPHK	Month-to-month	12	73.10	8.77	$100.15
4021-RQSNY	Month-to-month	12	67.90	8.14	$98.5
3793-MMFUH	Month-to-month	11	76.20	8.39	$95.05
0236-HFWSV	Month-to-month	11	74.20	8.16	$93.35

Figure 7. Top 10 Customer based on Retention Priority

---

d. Executive Summary

A single-page dashboard for leadership, summarizing KPIs and key drivers.

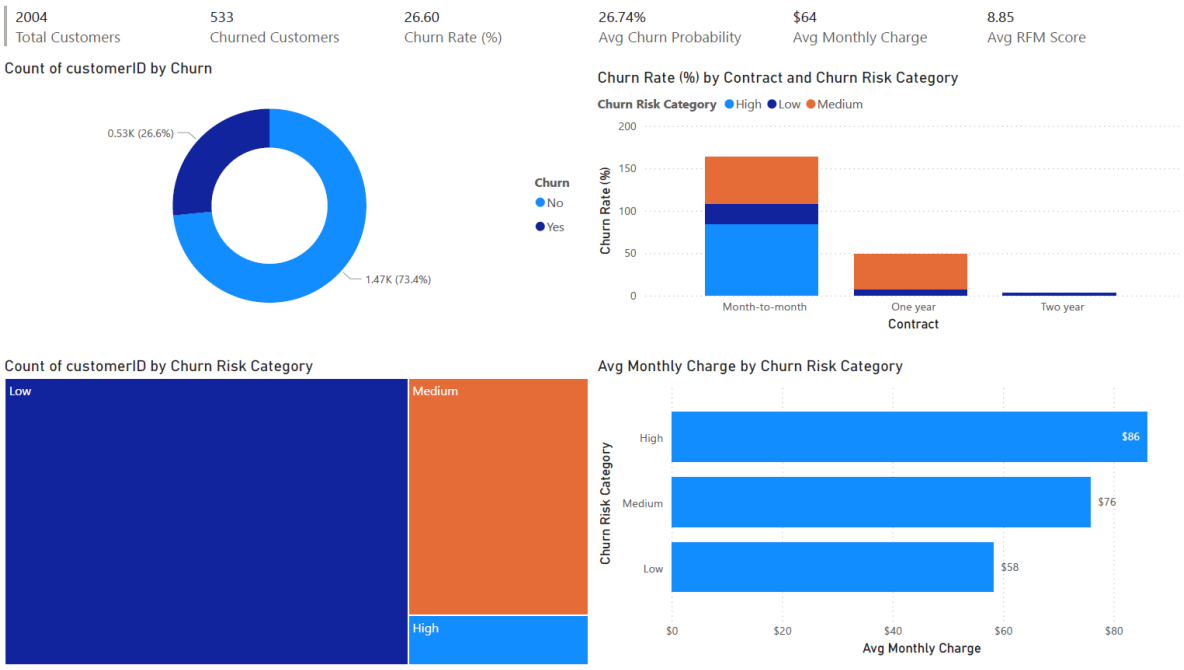

Figure 8 – Executive Summary Dashboard

---

4. Key Insights

Question	Finding
Who is likely to churn?	26.6 % customers predicted to churn — mostly Month-to-Month & Fiber users.
Why are they churning?	High monthly charges, short contracts, and payment issues.
Who is worth retaining?	107 customers with RFM > 8 and Churn Probability > 0.6.

---

5. Business Impact

- Improved retention targeting → potential 10–15 % reduction in churn.
- Identified $100K+ annual revenue at risk among high-value users.
- Enabled data-driven retention campaigns and measurable ROI.

---

6. Skills Demonstrated

Data Engineering: PySpark, Databricks FileStore, Feature Engineering
Machine Learning: Logistic Regression, Model Tuning, Evaluation (AUC, F1)
Data Visualization: Power BI (DAX Measures, Calculated Columns, Multi-Page Reports)
Business Analytics: RFM Scoring, Customer Segmentation, Churn Prediction

***