Customer Churn Prediction on the IBM Telco Dataset

Github: https://github.com/kasroma777/final-project
Prepared by Shamma Alhosani and Ksenia Romanova
December 16, 2025

Executive summary

This project predicts whether a telecom customer will churn (binary classification) using the IBM Telco Customer Churn dataset. The workflow is fully reproducible using scikit-learn pipelines for preprocessing and modeling. To preserve an unbiased final evaluation, the decision threshold is chosen on a validation set (not the test set), and then evaluated once on a held-out test set.

On the test set, the tuned Logistic Regression model achieves a ROC-AUC of 0.842, indicating strong probability ranking. Using a validation-chosen threshold of 0.35, the model achieves churn-class precision/recall/F1 of 0.561 / 0.717 / 0.629 (overall accuracy 0.776). A probability calibration check yields a Brier score of 0.1376. Additional analyses examine capacity-based targeting (top X% outreach) and subgroup performance across contract types and tenure bands.

1. Data description and preparation

The dataset contains 7,043 customers with a churn rate of approximately 26.5%. Features include customer demographics, service subscriptions (internet type, support/security add-ons, streaming services), and billing/contract information (contract length, payment method), along with numeric billing variables such as tenure, monthly charges, and total charges. The customerID identifier is removed prior to modeling.

TotalCharges cleaning and validation.

In the raw file, TotalCharges is stored as text and contains blank entries. After converting TotalCharges to numeric, 11 rows become missing. Before imputing, we validated that these rows correspond to new customers: 100% of missing TotalCharges rows have tenure == 0. Because total charges represent accumulated bill-to-date, imputing TotalCharges to 0.0 for tenure==0 customers is consistent with the business meaning of the variable and avoids dropping data.

Feature engineering.
Two lightweight, interpretable features are added: (1) tenure_bucket (0, 1-12, 13-24, 25-48, 49-72 months) to capture nonlinear tenure effects, and (2) high_monthly_charges indicating whether MonthlyCharges is at or above the 75th percentile cutoff (89.85).

2. Models and evaluation design

To prevent test-set leakage, the data is split into three partitions: train (60%), validation (20%), and test (20%), each stratified by churn. Preprocessing is implemented with a scikit-learn Pipeline + ColumnTransformer: numeric features are imputed (median) and standardized, and categorical features are imputed (most frequent) and one-hot encoded (handle_unknown='ignore').

We compare three models: a majority-class baseline (DummyClassifier), Logistic Regression (tuned with 5-fold cross-validation on the training split), and K-Nearest Neighbors (also tuned with 5-fold CV). Model selection uses ROC-AUC, which evaluates ranking quality across thresholds. The final decision threshold is selected on the validation split (either by maximizing F1 or to meet a capacity constraint) and then held fixed for one-time evaluation on the test split.

3. Results

3.1 Model comparison (default threshold = 0.50)

At the default 0.50 threshold, Logistic Regression achieves the highest ROC-AUC, indicating the strongest overall probability ranking. KNN is competitive and tends to trade slightly lower precision for higher recall. Because churn interventions are typically capacity- and cost-constrained, the operating point should be chosen explicitly rather than defaulting to 0.50.

Model	Test ROC-AUC	Accuracy	Churn precision	Churn recall	Churn F1
Baseline (most frequent)	0.500	0.735	0.000	0.000	0.000
Logistic Regression (tuned)	0.842	0.800	0.654	0.521	0.580
KNN (tuned)	0.830	0.786	0.599	0.583	0.591

3.2 Threshold selection without test-set leakage

Using validation predicted probabilities from the tuned Logistic Regression model, we select the threshold that maximizes F1. The chosen threshold is 0.35. Applying this frozen threshold to the test set yields churn precision/recall/F1 of 0.561 / 0.717 / 0.629 and overall accuracy 0.776 (ROC-AUC unchanged at 0.842).

At threshold 0.35, the test-set confusion matrix is: TN=825, FP=210, FN=106, TP=268. This operating point increases recall (identifying more churners) relative to the default threshold, at the cost of more false positives.

3.3 Capacity-based targeting (top X% outreach)

Many churn programs have a fixed outreach capacity (for example, a retention team can only contact the top 10% highest-risk customers). To support this use-case, we select a threshold on the validation probabilities that flags approximately the top X% of customers by risk, then evaluate that threshold on the test set.

Top X% targeted	Validation threshold	Test positive rate	Test accuracy	Precision	Recall	F1
5%	0.723	0.065	0.767	0.747	0.182	0.292
10%	0.652	0.109	0.787	0.740	0.305	0.432
15%	0.590	0.146	0.793	0.699	0.385	0.497
20%	0.516	0.202	0.800	0.662	0.503	0.571
30%	0.380	0.317	0.778	0.568	0.679	0.619

3.4 Calibration and subgroup checks

We assess probability calibration using the Brier score on the test set (lower is better). The Logistic Regression model achieves a Brier score of 0.1376, suggesting reasonably calibrated probabilities overall. We also evaluate whether performance differs across key subgroups at the final operating point (threshold 0.35).

Subgroup performance by contract type (threshold 0.35).

Contract	n	Churn rate	Precision	Recall	F1

Month-to-month	773	0.426	0.567	0.802	0.664
One year	300	0.120	0.333	0.111	0.167
Two year	336	0.027	0.000	0.000	0.000

Subgroup performance by tenure bucket (threshold 0.35).

Tenure bucket	n	Churn rate	Precision	Recall	F1
0	3	0.000	0.000	0.000	0.000
1-12	446	0.482	0.617	0.809	0.700
13-24	206	0.291	0.522	0.783	0.627
25-48	304	0.211	0.481	0.578	0.525
49-72	450	0.078	0.345	0.286	0.312

4. Interpretation and feature importance

Logistic Regression coefficients are useful for directional interpretation, but billing-related variables (MonthlyCharges, TotalCharges, tenure) are correlated (TotalCharges is roughly tenure * MonthlyCharges). This correlation can make individual coefficients less stable to interpret in isolation. To obtain a more robust view of what the fitted pipeline relies on, we compute permutation importance (scored by ROC-AUC) by permuting the original input columns and measuring performance drop.

Permutation importance (top features, test set, scoring = ROC-AUC).

Feature	Importance mean	Importance std
tenure	0.148100	0.004535
Contract	0.038514	0.004400
InternetService	0.033744	0.007271
MonthlyCharges	0.022218	0.004412

tenure_bucket	0.006163	0.001636
OnlineSecurity	0.003244	0.001190
StreamingMovies	0.002895	0.001751
TechSupport	0.002830	0.000973
StreamingTV	0.002759	0.002197
PaymentMethod	0.002562	0.001290

These results indicate that tenure and contract structure dominate the predictive signal, followed by internet service type and monthly charges. Support/security add-ons also contribute, consistent with EDA patterns that show lower churn among customers with services like online security and tech support.

5. Conclusion and next steps

This project demonstrates an end-to-end churn prediction workflow with an emphasis on reproducible preprocessing, careful evaluation design, and decision-focused reporting. Logistic Regression achieves strong probability ranking (test ROC-AUC 0.842) and, using a validation-selected threshold (0.35), provides a practical operating point for churn outreach with recall 0.717 and precision 0.561.

Next steps to further strengthen an operational churn system include: (1) cost-based thresholding using explicit assumptions about retention offer cost and expected retained value, (2) monitoring for data drift and performance degradation over time, and (3) a retraining schedule (for example, monthly or quarterly retraining depending on business cadence).

Data source

This project uses IBM's Telco customer churn sample dataset (a telecommunications company churn dataset) distributed with IBM Cognos Analytics. A commonly used public mirror is available on Kaggle as Telco Customer Churn (file: WA_Fn-UseC_-Telco-Customer-Churn.csv). Accessed December 15, 2025.