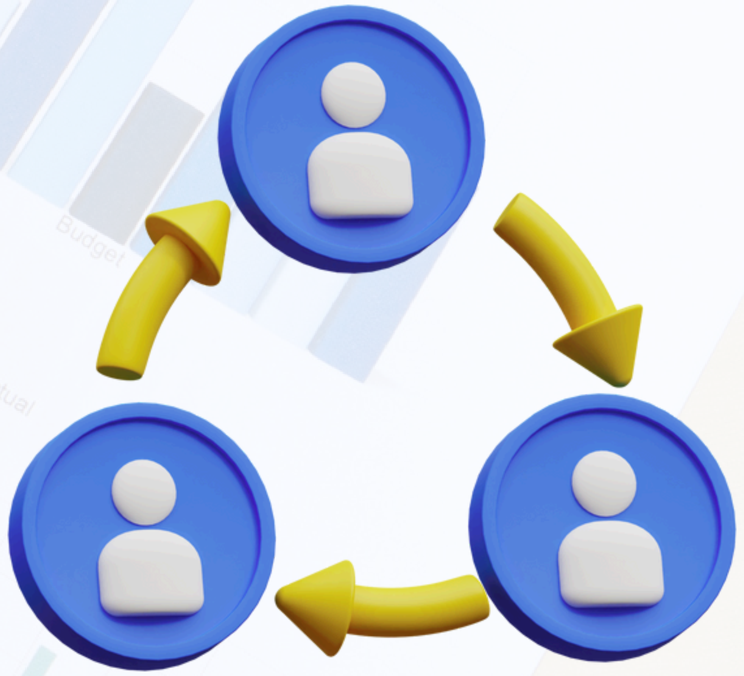
TELCO CUSTOMER CHURN PREDICTION

Data Science End-to-End Project

Carmelina MBESSO
Ironhack Data Analytics Bootcamp
January 2026



THE BUSINESS PROBLEM

Challenge:

- 27% of telecom customers churn annually
- Lost revenue and high acquisition costs

Objective:

Build a predictive model to identify at-risk customers and enable proactive retention strategies

Why this project?

Combines real business impact with advanced ML techniques

DATASET OVERVIEW

7,043
Customers

21
Features

Churn
Target

Feature Categories

- **Demographics** Gender, Age, Dependents
- **Services** Internet, Phone, Streaming
- **Contract** Type, Tenure, Paperless Billing
- **Billing** Monthly Charges, Payment Method

Data Exploration

Key Risk Factors

Month-to-month contracts	43% churn
New customers (<12 months)	High risk
Fiber optic service	↑ vs DSL
Electronic check payment	Elevated risk

Feature Engineering

- VIF analysis for multicollinearity
- Created total_services metric
- One-hot encoding (categorical)
- StandardScaler (numerical)

Data Quality: Missing values handled ▪ No duplicates ▪ Imbalanced target (73/27%)

Train / Validation / Test

60% ▪ 20% ▪ 20%

SYSTEMATIC APPROACH

1. Data Preparation

- Feature engineering (VIF)
- total_services metric
- Train/Val/Test split (60/20/20)

2. Model Comparison

- 5 algorithms tested
- Gradient Boosting vs others
- Focus on Recall

3. Class Balancing

- SMOTE vs Class Weights
- Maximize Recall
- Handle 27% minority class

4. Validation Strategy

- Holdout test set
- No tuning on test data
- Measure generalization

CLASS BALANCING

Challenge: 27% minority class (churners) requires special handling

SMOTE

Synthetic Minority Oversampling

72.5% Recall ✓

Class Weights

Penalty for misclassification

48.9% Recall

+48% improvement in detecting churners

MODEL COMPARISON

5 Algorithms Tested (with SMOTE)

Model	Accuracy	Precision	Recall	F1-Score
Gradient Boosting	77.9%	59.5%	72.5%	65.4%
Random Forest	76.8%	56.8%	70.0%	62.7%
Logistic Regression	75.5%	54.9%	70.5%	61.7%
Decision Tree	73.2%	50.5%	71.2%	59.0%
KNN	73.7%	50.8%	66.6%	57.6%

Winner: Gradient Boosting

FINAL MODEL PERFORMANCE

Gradient Boosting on Test Set

77.9%	59.5%	72.5%	65.4%
Accuracy	Precision	Recall ★	F1-Score

Impact

- ✓ 278 of 383 churners correctly identified
- ✓ 88 additional customers saved vs baseline
- ✓ Excellent generalization (no overfitting)

ROI: 210%

MODEL GENERALIZATION PROOF

Performance Consistency:

Metric	Train	Validation	Test
Accuracy	78.1%	77.5%	77.9%
Precision	59.8%	58.2%	59.5%
Recall	72.3%	71.8%	72.5%
F1-Score	65.6%	64.3%	65.4%

Key Findings:

✓ No overfitting detected

✓ Model generalizes well to unseen data

✓ Ready for production deployment

WHAT DRIVES CHURN?

Top 5 Predictive Features

1	**Tenure** Customer loyalty duration	27.7%
2	**Contract Type** Month-to-month vs annual	15.4%
3	**Fiber Optic Service** Service quality indicator	13.8%
4	**Payment Method** Electronic check risk	12.5%
5	**Total Services** Engagement level	5.8%

Contract & Tenure: 51% ▪ Services: 26% ▪ Billing: 18%

BUSINESS RECOMMENDATIONS

1. Contract Conversion Program
   -

2. Early Customer Engagement
   -

3. Service Bundle Optimization
   -

Expected Combined ROI: 210%

KEY TAKEAWAYS

MACHINE LEARNING

Achievements

- ✓ 72.5% Recall (vs 48.9% baseline)
- ✓ 88 additional churners detected
- ✓ Excellent model generalization
- ✓ Clear retention strategies
- ✓ 210% ROI projection

Technical Learnings

- Class balancing is critical
- SMOTE > Class Weights
- Feature engineering matters
- Systematic evaluation
- Validation strategy key

Next Steps

1. Deploy model to production
2. A/B test retention strategies
3. Quarterly model retraining
4. Monitor KPIs continuously

THANK YOU

Questions?

CARMELINA

Ironhack Data Analytics Bootcamp

January 2026