Datasheet

for Customer Churn Prediction Model

# 1. Motivation

## – Purpose

The dataset was released by IBM Sample Analytics to illustrate customer-churn modelling for subscription-based businesses. It is widely used for teaching, benchmarking classical ML pipelines, and rapid prototyping of binary-classification methods.

## – Supporting tasks

Primary: Predict whether a telco customer will churn in the next period (binary classification).

Secondary: Exploratory data analysis on usage patterns, survival/retention modelling, uplift modelling, cost-sensitive learning, explainability demos.

## – Creators & funders

• Original creator: IBM Sample Analytics Team (publicly posted on IBM Community & Kaggle).

• Funding: Internal IBM resources (no outside grant reported).

## – Beneficial uses

– Educational tutorials on feature engineering, class imbalance, model evaluation.

– Benchmarking lightweight tabular AutoML systems.

– Demonstrations of fairness/explainability tooling.

## – Misuse and out-of-scope uses

– Re-identification or linkage attacks on hypothetical individuals.

– Drawing real-world demographic conclusions (gender, senior-citizen) about actual telco populations—the data are synthetic/obfuscated.

– Deployment in production without extensive validation—field distributions may differ.

# 2. Composition

## – Instances

7 043 customer records, each representing a unique billing account at a given snapshot in time.

– Features (21)

Numerical (2)

 • tenure – number of months with the company (int)

 • MonthlyCharges – current monthly fee (float)

 • TotalCharges – life-time spend (float)

Categorical/Binary (18)

 customerID, gender, SeniorCitizen, Partner, Dependents, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, Churn (target variable).

## – Missing data

• 11 rows have blank “TotalCharges” where tenure = 0.

## – Sensitive attributes

• gender, age proxy (SeniorCitizen) could be used as sensitive or legally protected classes.

## – Class balance

• Churn = “Yes”: 26.5 % (1 869)

• Churn = “No”: 73.5 % (5 174)

## – Is the data synthetic?

Yes—values are simulated/perturbed; no personally identifiable information remains.

# 3. Collection process

## – Original raw source

Generated by IBM Analytics from an undisclosed U.S. telco-like schema; released via IBM community blog c. 2018.

## – Sampling strategy

• Undisclosed. Likely convenience sampling to match realistic churn rates.

## – Timeframe

• Snapshot; tenure indicates up to 72 months of history but no explicit date field.

## – Consent & Privacy

• Since data are synthetic, no direct consent needed.

# 4. Pre-processing / cleaning / labeling

## – Pre-processing done by original release

• Categorical values coerced to strings.

• SeniorCitizen stored as 0/1 ints (not booleans).

• customerID is a random hash.

• TotalCharges appears as a string; must be converted to float with coercion.

## – Your recommended additional steps

1. Strip whitespace and cast TotalCharges to numeric; drop 11 NA rows or impute 0.

2. Encode categoricals (one-hot or target encoding).

3. Check multicollinearity (TotalCharges highly correlated with tenure × MonthlyCharges).

4. Address class imbalance (e.g. stratified splits, weighting, SMOTE).

5. Consider fairness auditing for gender & SeniorCitizen.

# 5. Distribution

## – License

• Creative Commons CC0 1.0 Universal (public domain) per IBM sample-data policy (confirmed in repository).

## – How to obtain

• File link above or via Kaggle “Telco-Customer-Churn” dataset.

## – Citation

IBM Sample Data Sets: Telco Customer Churn. IBM Developer Community (2018).

# 6. Uses

## – Training / Validation splits

• Typical 70 / 30 stratified or k-fold cross-validation.

• Time-based split impossible (no timestamp).

## – Performance metrics

• Accuracy, ROC-AUC, PR-AUC, recall at fixed precision (because churn is minority).

• Financial cost metrics (e.g. expected monthly revenue saved).

## – Fairness & ethics testing

• Evaluate disparate impact across gender & SeniorCitizen.

• Check calibration curves per subgroup.

# 7. Maintenance

## – Owner of the canonical copy

• IBM Sample Analytics GitHub.

## – Updates

• No planned updates—static dataset.

## – Bug reporting

• Open an issue in the IBM sample-data repository or email team@julius.ai with replication details.

# 8. Ethical considerations & biases

## – Representational risks

• Gender variable only contains “Male” and “Female” → non-inclusive.

• SeniorCitizen is binary thresholded at 65 yrs (assumed); discards nuanced age info.

## – Statistical biases

• Synthetic generation may not reflect real churn drivers; models trained here risk poor external validity.

## – Societal & legal risks

• If deployed uncautiously, churn-prediction systems can be used for differential pricing that discriminates against protected groups.

## – Risk mitigation recommendations

– Document model purpose and protected-attribute impacts during evaluation.

– Apply fairness constraints or post-processing if deploying in sensitive domains.

– Communicate that the dataset is synthetic and not representative of any country’s demographics.

# 9. Relevant papers / benchmarks

– Predicting Customer Churn – A Case Study in Telecommunication Service Providers.

– AutoGluon, H2O.ai AutoML, TPOT frequently benchmark on this dataset.

# 10. Version history

v1 (2018) – Original IBM release