Causal Impact of Contract Type on Customer Churn

University of Southern California
DSO 599 Special Topic: Causal Inference with Machine Learning for Business Analytics
Amna Gul, Anita Lin, Joey Cindass, Roxanne Li
December 2025

Overview

Customer churn is a central challenge in subscription-based industries, where firm profitability depends heavily on retaining existing customers rather than acquiring new ones. In the telecommunications sector, contract type, specifically whether a customer is enrolled in a month-to-month or long-term (one- or two-year) contract, represents a policy-relevant and directly manipulable intervention that may influence churn behavior through pricing, commitment, and switching costs. This paper studies the causal effect of long-term contract enrollment on a customer's probability of churn using observational customer-level data. We define treatment as enrollment in a long-term contract and the outcome as a binary indicator of churn at the individual customer level. Because contract choice is endogenous and driven by observed and unobserved customer characteristics such as tenure, service usage, and billing behavior, naïve comparisons of churn rates across contract types are confounded and not causally interpretable. To address this selection problem, we employ propensity score–based methods and doubly robust estimation to approximate a randomized experiment under explicit identification assumptions. The resulting estimates aim to isolate the causal impact of contract length on churn and provide guidance for contract design and customer retention strategies.

1. Causal Question and Motivation

1.1 Causal Question and Estimand

This study investigates the causal effect of contract type on customer churn in the telecom industry. Specifically, we ask:

What is the causal effect of being on a long-term contract (one- or two-year) versus a month-to-month contract on a customer's probability of churning?

Let $Y(1)$denote the potential churn outcome if a customer is on a long-term contract, and $Y(0)$ the potential outcome if the same customer is on a month-to-month contract. Our target estimand is the Average Treatment Effect (ATE):

$$ATE = E[Y(1) - Y(0)]$$

The treatment indicator is defined as:

- $D = 1$: customer is on a long-term contract (one-year or two-year),
- $D = 0$: customer is on a month-to-month contract.

1.2 Motivation

Telecom firms operate in highly competitive, subscription-based markets where profitability depends critically on retaining existing customers. Churn directly reduces recurring revenue and forces firms to incur acquisition and marketing costs to replace lost customers. Even modest reductions in churn can translate into substantial gains in customer lifetime value (CLV).

Contract type is a key lever that telecom firms can actively design and manage. Long-term contracts may reduce churn through price discounts, switching frictions, and stronger commitment; however, customers self-select into contract types based on unobserved preferences and risk. Raw churn differences therefore conflate the true causal effect of contract type with underlying differences in customer characteristics.

Estimating the causal effect of contract type is thus directly relevant for:

- Designing and pricing long-term contract offers;
- Optimizing retention and upsell strategies for month-to-month customers;
- Quantifying the incremental CLV impact of encouraging contract migration.

Recent research and industry practice emphasize moving from predictive churn models to causally interpretable estimates of intervention effects, such as contract changes, to guide strategic decisions.

2. Data

2.1 Dataset Overview

We use the Telco Customer Churn dataset (2019), a cross-sectional snapshot of 7,043 unique customers' records containing:

- Contract information: month-to-month, one-year, or two-year contracts.
- Churn status: whether the customer discontinued service.
- Demographics: gender, senior citizen indicator, partner, dependents.
- Service features: phone service, internet service, streaming services, security add-ons.
- Billing and payment: monthly charges, total charges, electronic billing, payment method.
- Tenure: number of months with the company.

Each row corresponds to a unique customer. There are no repeated measurements over time.

2.2 Treatment, Outcome, and Covariates

- Treatment D:
    - D=1 if the customer is on a one-year or two-year contract;

    - D=0 if the customer is on a month-to-month contract.
- Outcome Y:
    - Y=1 if the customer churned
    - Y=0 if customer has not churned
- Covariates X:
  All demographic, billing, service, and tenure variables—used to adjust for confounding.

2.3 Cleaning and Preprocessing

Key preprocessing steps include:

- Missing data: The TotalCharges variable had 11 missing values, all associated with tenure = 0. These entries were set to 0 and the variable converted to numeric.
- Binary encoding: All Yes/No variables were converted to 1/0.
- Categorical Variables: Contract type, InternetService, and PaymentMethod were one-hot encoded with drop_first=True to avoid multicollinearity.
- Service categories (e.g. No Internet service / No Phone service): Encoded as 0.
- CustomerID: Removed because it contains no information about causality.

After preprocessing, the dataset is fully numeric and ready for causal inference methods.

2.4 Descriptive Statistics

Descriptive patterns indicate substantial differences across contract types:

- Approximately 45% of customers are on long-term contracts, with an average churn rate of about 11%. (*see Figure 2.4.1*)
- Approximately 55% are on month-to-month contracts, with an average churn rate of about 43%. (*see Figure 2.4.2*)
- The overall churn rate in the sample is approximately 26.5%. (*see Figure 2.4.3*)

These sharp raw differences motivate the use of causal inference methods to separate selection effects from the genuine impact of contract length.

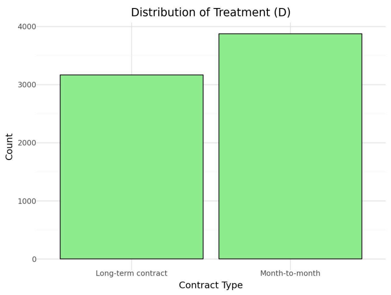


Figure 2.4.1

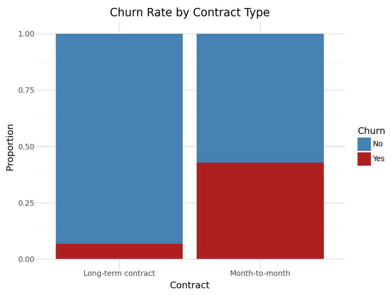


Figure 2.4.2

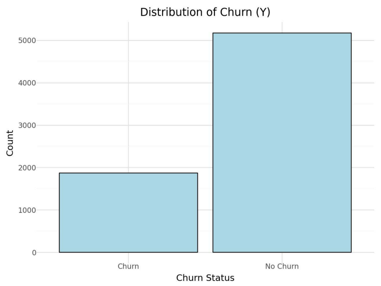


Figure 2.4.3

3. Methods

3.1 Identification Strategy

Because contract type is not randomly assigned, simple comparisons of churn rates across contract types will be biased. We adopt a propensity-score-based identification strategy, using:

- Inverse Probability Weighting (IPW) to reweight observations and approximate a randomized experiment;
- A Doubly Robust (DR) estimator, combining IPW with outcome regression to improve robustness.

Our identification relies on the following assumptions:

1. **Unconfoundedness (Selection on Observables):**
   Conditional on observed covariates (demographics, tenure, pricing, and service mix), assignment to long-term vs. month-to-month contracts is assumed to be as good as random. This is reasonable because these factors strongly influence real-world contract choices. However, unobserved factors—such as satisfaction, desire for flexibility, or competitive offers—may still affect both contract choice and churn. If more satisfied or risk-averse customers both choose long-term contracts and churn less, our estimates could overstate the true effect.
2. **Overlap (Positivity):**
   For most covariate profiles, customers have a positive probability of being in either contract type. Overlap may fail for very new customers or those with specialized service bundles who almost always choose one contract type.
3. **Stable Unit Treatment Value Assumption (SUTVA):**
   Each customer's churn outcome depends only on their own contract type, not on other customers' contracts. This is plausible because churn is generally driven by individual service quality and pricing.
   Minor violations could occur if large contract promotions affect network performance, but such spillovers are likely small.
4. **Correct Specification (for DR):**
   For the doubly robust estimator, consistency requires that at least one of the following be correctly specified:
   - The propensity score model $e(x)$, or
   - The outcome models $m_1 = E[Y|D = 1, X]$ and $m_0 = E[Y|D = 0, X]$

3.2 Propensity Score Estimation

We estimate the propensity score, $e(X) = P(D = 1|X)$, using models based solely on pre-treatment covariates X. To mitigate the influence of extreme probabilities on the IPW estimator, we clip the estimated scores to the interval $[0.01, 0.09]$.

As a robustness check, we also estimated the propensity score with a random forest model. While the random forest produced *more extreme* propensity scores and worse overlap/balance (heavier mass near 0 and 1), the resulting IPW ATE was very similar in magnitude and sign to the logistic regression–based estimate. Because logistic regression achieves better covariate balance and more reasonable overlap while delivering nearly identical treatment effects, we use it as our primary propensity score specification, and view the tree-based results as supporting evidence that our main conclusions are not driven by the specific propensity score model. (see section **3.5.3 Alternative Propensity Score Model: Random Forest Diagnostics** for more information)

3.3 Inverse Probability Weighting (IPW)

Given the estimated propensity scores, we define the IPW weights:

- For treated units (D=1): $w_i = \frac{1}{e(X_i)}$
- For control units (D=0): $w_i = \frac{1}{1-e(X_i)}$

IPW reweights the sample so that, in expectation, the distribution of observed covariates among treated and control units becomes similar, approximating the balance that would be achieved under randomization. The ATE is then estimated by the weighted difference in average outcomes between the two groups.

3.4 Doubly Robust (DR) Estimation

To further guard against misspecification, we implement a doubly robust estimator. We estimate two conditional outcome models:

- $m_1(X) = E[Y|D = 1, X]$,
- $m_0(X) = E[Y|D = 0, X]$,

and then form a DR estimator that combines these models with IPW-based correction terms. The key property of the doubly robust estimator is that it remains consistent if either the propensity score model or the outcome models are correctly specified, but not necessarily both.

3.5 Diagnostic Checks

3.5.1 Propensity Score (Estimated by Logistic Regression) Overlap

We inspect the distribution of estimated propensity scores using logistic regression by treatment status. Long-term customers tend to have high propensity scores, whereas month-to-month customers often have low propensity scores. Although clipping reduces the impact of extremes, the distributions still show limited overlap near the tails, indicating that common support is imperfect, especially for very low and very high propensity values. (see Figure 3.5.1)

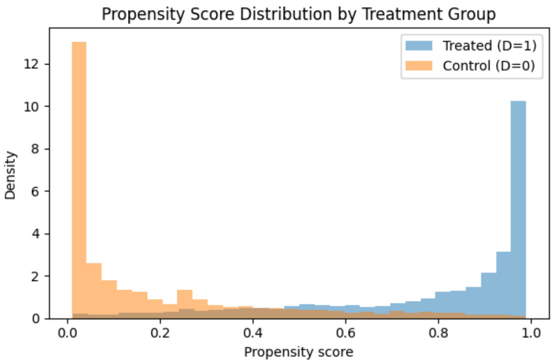


Figure 3.5.1

3.5.2 Covariate Balance

To evaluate how well IPW balances the covariates, we compute standardized mean differences (SMDs) for each covariate before and after weighting. After applying IPW:

- Most covariates exhibit |SMD| < 0.1, a common threshold for acceptable balance.
- However, tenure and TotalCharges remain moderately imbalanced, with |SMD| ≈ 0.37 and 0.25, respectively.

These diagnostics indicate that IPW greatly improves balance but does not fully eliminate differences in time-related variables. (see Figure 3.5.2)

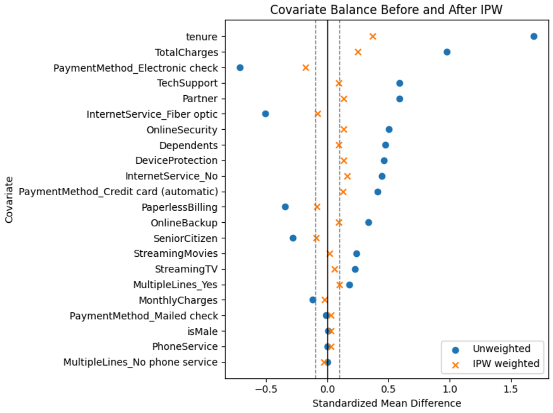


Figure 3.5.2

3.5.3 Alternative Propensity Score Model: Random Forest Diagnostics

As a robustness check, we re-estimated the propensity scores using a random forest classifier instead of logistic regression. The goal was to assess whether a more flexible, nonparametric model would improve overlap or balance.

The diagnostics show that the random forest produced more extreme propensity scores, with much of the control group concentrated near 0 and the treated group near 1 (see Figure 3.5.3a). This led to worse covariate overlap compared to the logistic model and larger portions of the sample with effectively no comparable units across treatment groups. Correspondingly, the covariate balance after IPW weighting was inferior to the balance achieved using the logistic regression propensity scores (Figure 3.5.3b).

Despite these differences in overlap and balance, the resulting IPW ATE estimate was very similar in magnitude and sign to the logistic regression–based estimate. This indicates that our main conclusions are not sensitive to the specific choice of propensity score model, although the logistic regression specification remains preferable due to its superior overlap and balancing properties.

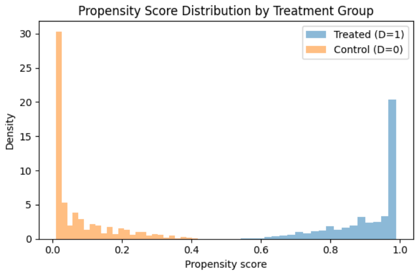


Figure 3.5.3a

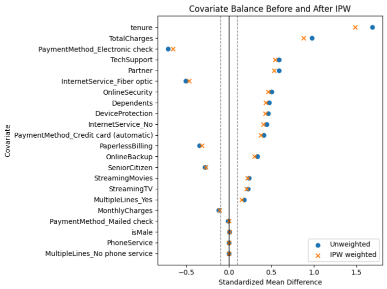


Figure 3.5.5b

3.5.4 Measures of Uncertainty

For both IPW and DR estimators, we quantify uncertainty using bootstrap standard errors and 95% confidence intervals, based on repeated resampling of the dataset.

4. Results

4.1 IPW Estimates

Using the IPW estimator, we obtain the following estimate of the average treatment effect:

$$\widehat{ATE}_{IPW} =- 0.2102$$

This implies that, on average, being on a long-term contract reduces the probability of churn by 21.0 percentage points relative to a month-to-month contract.
The associated uncertainty measures are:

- Bootstrap standard error: 0.0057
- 95% confidence interval: $[- 0.2215, \ - 0.1990]$

The IPW estimate is both statistically significant and economically large.

4.2 Doubly Robust (DR) Estimates

The doubly robust estimator yields a somewhat more conservative effect size:

$$\widehat{ATE}_{DR} =- 0.165$$

This indicates that long-term contracts reduce churn probability by 16.5 percentage points on average.
Uncertainty measures:

- Bootstrap standard error: 0.0151
- 95% confidence interval: $[-0.1947, -0.1353]$

The DR estimate remains strongly negative and statistically significant, but with a smaller magnitude than the IPW estimate, consistent with the additional adjustment provided by outcome modeling.

4.3 Other Methods Considered

We also experimented with Double Lasso and Double ML in our methodology, but these modifications did not significantly alter the results. (see code for details)

5. Robustness Check

5.1 Trimming for Improved Common Support

Because IPW estimators are sensitive to extreme propensity scores—where treated and control units have few comparable counterparts—we conducted a robustness check based on propensity score trimming. Specifically, we restricted the analysis to customers whose estimated propensity scores fall within [0.05, 0.95].

This procedure removes observations in the extreme tails where P(D=1|X) is nearly 0 or nearly 1, thereby focusing inference on a population with meaningful overlap between contract types. The resulting estimand corresponds to the Average Treatment Effect on the overlap population, a subset for which credible counterfactuals exist for both treatment states.

Figure 5.1.1 shows the weighted propensity score distributions after trimming to [0.05, 0.95]. Treated and control groups now overlap across the full range, and their weighted densities are closely aligned, indicating much better common support and less reliance on extrapolation for our causal comparisons.

Table 5.1.2 reports standardized mean differences (SMDs) before weighting (SMD_unw) and after IPW weighting (SMD_w) on the trimmed sample. Balance improves sharply: for example, tenure falls from 0.8238 to 0.0512, and TotalCharges from 0.5083 to 0.0237, with nearly all covariates below $|SMD| < 0.06|$. Overall, trimming plus IPW weighting yields excellent balance, making treated and control groups highly comparable on observed characteristics.

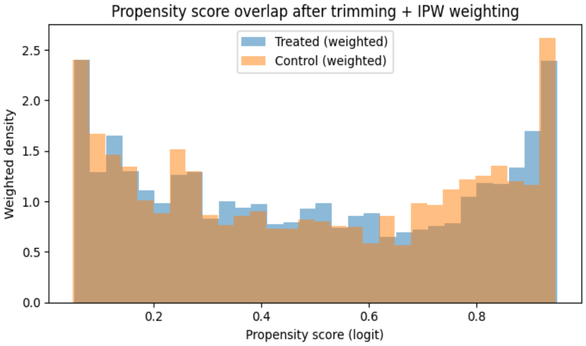
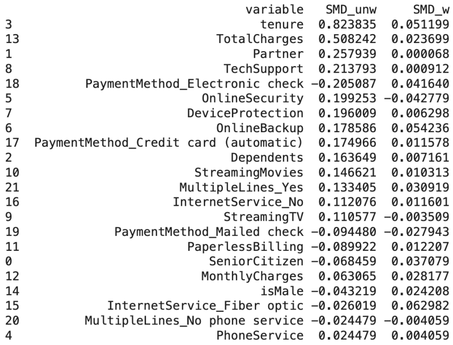


Figure 5.1.1



```
                                           variable   SMD_unw     SMD_w
3                                            tenure  0.823835  0.051199
13                                     TotalCharges  0.508242  0.023699
1                                           Partner  0.257939  0.000068
8                                       TechSupport  0.213793  0.000912
18                   PaymentMethod_Electronic check -0.205087  0.041640
5                                    OnlineSecurity  0.199253 -0.042779
7                                  DeviceProtection  0.196009  0.006298
6                                      OnlineBackup  0.178586  0.054236
17            PaymentMethod_Credit card (automatic)  0.174966  0.011578
2                                        Dependents  0.163649  0.007161
10                                  StreamingMovies  0.146621  0.010313
21                                MultipleLines_Yes  0.133405  0.030919
16                               InternetService_No  0.112076  0.011601
9                                       StreamingTV  0.110577 -0.003509
19                       PaymentMethod_Mailed check -0.094480 -0.027943
11                                 PaperlessBilling -0.089922  0.012207
0                                     SeniorCitizen -0.068459  0.037079
12                                   MonthlyCharges  0.063065  0.028177
14                                           isMale -0.043219  0.024208
15                      InternetService_Fiber optic -0.026019  0.062982
20                   MultipleLines_No phone service -0.024479 -0.004059
4                                      PhoneService  0.024479  0.004059
```

Figure 5.1.2

5.2 Result

After trimming and re-estimating the IPW model, we obtain:

- Trimmed IPW ATE:

  $\widehat{ATE_{IPW,\, trimmed}} = -\ 0.16$

- Bootstrap Standard Error (500 replications): 0.0146
- 95% CI (normal approximation): [−0.1442,−0.0870]

Even after restricting the analysis to customers with strong overlap and excellent covariate balance, the estimated causal effect remains negative and statistically significant, and

The magnitude (≈ 11.6 percentage points) is smaller than the full-sample IPW estimate (−21 p.p.), but this is expected: trimming removes extreme-propensity customers who heavily influence the original estimate and focuses instead on the realistic target population where treated and control customers are observationally comparable.

6. Discussion

6.1 Summary of Findings

Under plausible identifying assumptions and after adjusting for observed covariates, we find that:

- Long-term contracts causally reduce customer churn by approximately 16–21 percentage points compared with month-to-month contracts.
- This effect is economically meaningful in a subscription-based business, where small changes in churn can generate substantial differences in customer lifetime value.

These results confirm that the lower churn rates observed among long-term contract customers are not solely due to selection; the contract structure itself exerts a significant causal effect.

6.2 Limitations and Remaining Biases

Our strategy relies on the unconfoundedness assumption where all variables affect both contract choice and churn in which they are included in the model. Although this dataset does contain very important features, it does not capture more nuanced information such as customer satisfaction, the income of the customer, or relative competitor offers at the time of churn. Because it is impossible to say we could include all potential factors, our estimates may still be biased.

We must also consider that this data represents a single cross-sectional snapshot, which may limit findings related to temporal ordering. There are some features in the data that may be influenced by previous experiences, such as Tenure or Total Charges. While these features are informative, neglecting their temporal nature may introduce unwanted bias.

Our diagnostic checks show that there is an imperfect overlap within the propensity scores of contract types, especially at the extreme values. Although trimming improved the balance and robustness, there could still be some hidden unaddressed bias.

We assume that a customer’s contract type does not affect another customer’s churn outcome. But in reality, there could be hidden layers that could create spillovers that may violate the SUTVA assumption thus increasing bias.

6.3 Implications for Decision Makers

Long-term contracts should be used as a targeted retention tool and not just a blanket solution to reduce churn among all customer categories. However, decision makers should rely on our causal estimates and not just the descriptive churn rates when designing contract initiatives and allocating budgets. To maximize impact, decision makers should combine contract strategies - with regard to causal findings - with broader customer value and experience.

6.4 Directions for Future Work

Future work could strengthen these findings by leveraging panel data that track customers over time, allowing for more customer designs that may better establish causality. Incorporating satisfaction scores or usage could further strengthen causality, while exploring other treatment effects could reveal which customer segments benefit most from long term contracts, enabling more targeted strategies. Finally, looking further into other possible controllable levers, such as pricing, or service bundles, could help identify other causal factors that influence churn.

7. Reference

- Barsotti, A., Castellani, A., Musumeci, F., & Levorato, G. (2024). *A decade of churn prediction techniques in the TelCo sector: A systematic literature review.* Information Systems Frontiers. https://doi.org/10.1007/s10796-024-10573-6
- Devriendt, F., Verbeke, W., & Baesens, B. (2021). Why you should stop predicting customer churn and start predicting customer churners who can be retained. *Information Sciences, 572*, 522–539. https://doi.org/10.1016/j.ins.2021.05.002
- Gutierrez, P., & Gérardy, J.-Y. (2017). *Causal inference and uplift modeling: A review of the literature.* Proceedings of the PMLR. https://proceedings.mlr.press/v68/gutierrez17a.html
- Telco Customer Churn Analysis. (2022). *Measuring the effect of different contracts using causal inference.* Retrieved from Towards Data Science.
- Verhelst, T., Caelen, O., Dewitte, J.-C., Lebichot, B., & Bontempi, G. (2018). *Understanding telecom customer churn with machine learning: From prediction to causal inference.* Orange Belgium Research Report.
- Verhelst, T., Mercier, D., Shrestha, J., & Bontempi, G. (2023). *A churn prediction dataset from the telecom sector: A new benchmark for uplift modeling.* arXiv preprint. https://doi.org/10.48550/arXiv.2307.02435

8. Appendix

All methods described in the report—including IPW, DR estimation, balance diagnostics, and robustness checks—are implemented in the Python code provided below. Link:https://github.com/Guluna/Causal-Inference-Customer-Churn/blob/main/Causal_Impact_of_Contract_Type_on_Customer_Churn.ipynb