1. Data Exploration

a. Overview

This report analyzes the relationship between various customer attributes and their respective churn rates. The objective is to identify high-risk customer segments and uncover behavioral or demographic patterns that contribute to churn, thereby enabling the development of targeted retention strategies.

b. Key Insights

Demographic Factors

Age-related churn differences are modest but consistent, with rates ranging from about 14% to 15.6%. Customers aged 34–41 show the highest churn tendency, while those under 34 have slightly lower rates. Gender differences are small but steady: men churn at about 15.6% compared to 14.4% for women, suggesting a slight male predisposition to leave.

Marital and Family Status

Married customers churn at the highest rate (15.5%), followed by single individuals, while divorced customers show the lowest rates. Family size influences churn as well—customers with fewer than two dependents churn more often than those with larger families. This suggests that broader household responsibilities may strengthen loyalty to insurance policies.

Socioeconomic Indicators

Income and employment status reveal sharper churn contrasts. Low-income customers have the highest churn rate at nearly 16%, while higher earners stay more frequently. Retirees and unemployed customers also churn at higher rates compared to those who are employed, underscoring the stabilizing role of steady income. Education shows a less linear trend, with Master's degree holders at 15.5% churn and PhD holders at the lowest (14.5%).

Engagement Behavior

Customer interaction levels are among the strongest churn predictors. Those logging in once or less per month churn at over 28%, whereas customers logging in three or more times monthly are nearly fully retained. Mobile app users have a churn rate of just 7%, compared to over 20% for non-users. Email engagement is even more telling—open rates below 30% correspond to a 58% churn rate, highlighting disengagement as a key risk factor.

Customer Service Experience

Service-related friction has a substantial impact. Customers making more than two service calls per year churn at a staggering 62%, suggesting that unresolved issues or repeated complaints drive dissatisfaction. Proactive problem resolution could significantly improve retention in this group.

Policy and Payment Features

Policy characteristics also play a role. Term life products and lower coverage amounts tend to be linked to higher churn, whereas universal life policies see better retention. Annual premium payment customers churn at the lowest rate (12.6%), outperforming monthly payers at 15.4%. Automatic payment methods also slightly improve retention over manual payments.

Recency and Activity Gaps

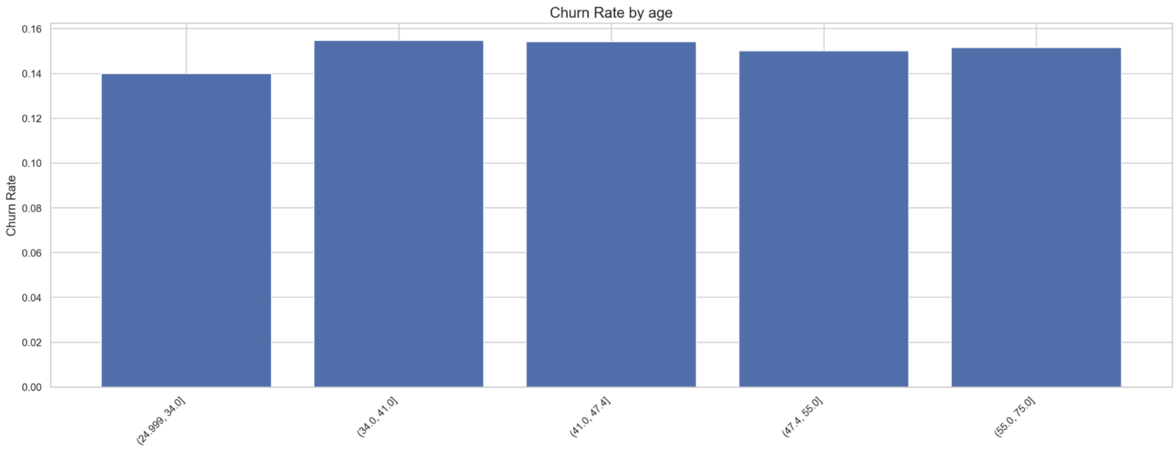
The number of days since last login is another meaningful signal—customers inactive for over 341 days show an uptick in churn risk. This reinforces the importance of re-engagement campaigns and periodic check-ins. Metrics like days between policy start and churn reflect post-churn behavior more than serve as predictive inputs but still help contextualize timelines.

Overall

Churn risk is driven most strongly by low engagement, poor service experiences, and financial strain. Retention strategies should focus on increasing digital interactions, improving service resolution, and offering flexible options for cost-sensitive customers to improve long-term loyalty.

c. Feature Enginnering

	value	churn_rate
0	(24.999, 34.0]	0.139912
1	(34.0, 41.0]	0.154744
2	(41.0, 47.4]	0.154116
3	(47.4, 55.0]	0.150068
4	(55.0, 75.0]	0.151582



This chart is part of the Single-Segment Churn Patterns view, which allows you to explore churn behavior for one customer attribute at a time using the dropdown selector at the top. In this case, the chosen feature is age. The system has automatically divided customers into five equal-sized age ranges (quintiles) and calculated the churn rate for each. These ranges appear in the first column of the table, while the second column shows the proportion of customers within each range who have churned.

The bar chart below the table provides a visual comparison of these churn rates across the age groups. In this dataset, customers aged 25–34 have the lowest churn rate at around 14%. The remaining groups — 34–41, 41–47.4, 47.4–55, and 55–75 — all show churn rates clustered between 15.0% and 15.5%. This narrow range suggests that age alone is not a strong driver of churn, as the variation across groups is minimal.

While age appears to have only a weak correlation with churn, the dropdown menu allows you to select other features, such as income bracket, policy type, or login frequency, which may reveal more pronounced differences. This makes the tool useful for quickly scanning through variables to identify those with greater potential impact on churn.

Cross-Segment Churn

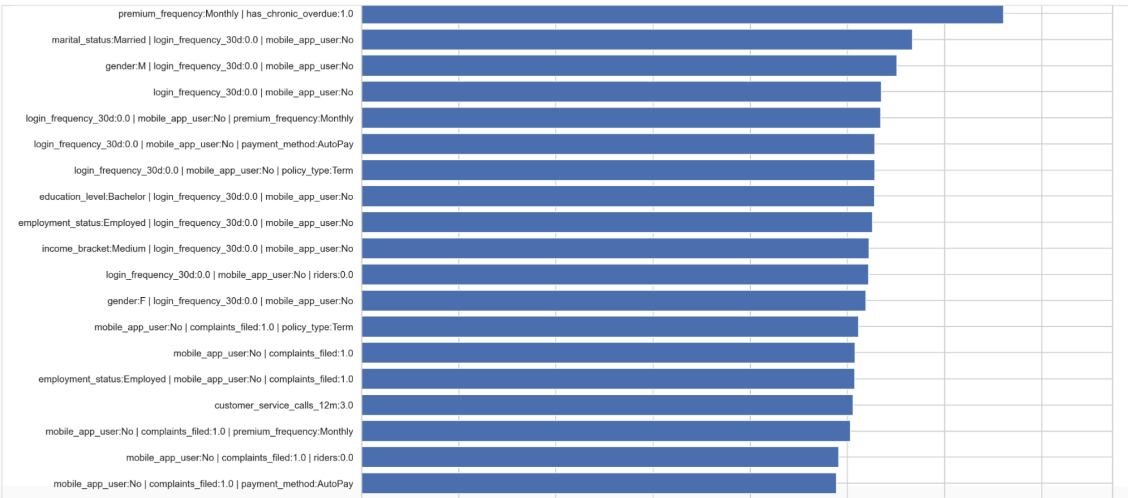
Min sample... 500 Max comb... 3 Run

```
Top 30 High-Churn Cross Segments (min samples = 500)
```

	variables	segment_values	count	churn_rate
65	engagement_score	engagement_score:-1.0	644	0.736025
778	mobile_app_user, engagement_score	mobile_app_user:No | engagement_score:-1.0	624	0.727564
1132	has_chronic_overdue, engagement_score	has_chronic_overdue:0.0 | engagement_score:-1.0	519	0.682081
6447	mobile_app_user, has_chronic_overdue, engageme...	mobile_app_user:No | has_chronic_overdue:0.0 |...	501	0.670659
64	has_chronic_overdue	has_chronic_overdue:1.0	684	0.665205
1004	premium_frequency, has_chronic_overdue	premium_frequency:Monthly | has_chronic_overdu...	663	0.660633
2541	marital_status, login_frequency_30d, mobile_ap...	marital_status:Married | login_frequency_30d:0...	769	0.566970
1659	gender, login_frequency_30d, mobile_app_user	gender:M | login_frequency_30d:0.0 | mobile_ap...	710	0.550704
644	login_frequency_30d, mobile_app_user	login_frequency_30d:0.0 | mobile_app_user:No	1398	0.535050
5620	login_frequency_30d, mobile_app_user, premium_...	login_frequency_30d:0.0 | mobile_app_user:No |...	998	0.534068
5627	login_frequency_30d, mobile_app_user, payment_...	login_frequency_30d:0.0 | mobile_app_user:No |...	922	0.528200
5614	login_frequency_30d, mobile_app_user, policy_type	login_frequency_30d:0.0 | mobile_app_user:No |...	852	0.528169
5078	education_level, login_frequency_30d, mobile_a...	education_level:Bachelor | login_frequency_30d...	614	0.527687
4560	employment_status, login_frequency_30d, mobile...	employment_status:Employed | login_frequency_3...	909	0.525853
3924	income_bracket, login_frequency_30d, mobile_ap...	income_bracket:Medium | login_frequency_30d:0....	651	0.522273
5634	login_frequency_30d, mobile_app_user, riders	login_frequency_30d:0.0 | mobile_app_user:No |...	849	0.521790
1654	gender, login_frequency_30d, mobile_app_user	gender:F | login_frequency_30d:0.0 | mobile_ap...	688	0.518895

Cross-Segment Churn

Min sample... 500 Max comb... 3 Run



This chart presents the Top 30 High-Churn Cross Segments for customers with at least 500 records in each segment. Unlike single-variable churn analysis, which looks at one feature at a time, cross-segment analysis examines combinations of variables to identify customer groups where churn is disproportionately high. Each bar represents a specific combination of attributes, with its length corresponding to the churn rate for that group.

At the top of the chart, the most churn-prone segment consists of customers with an engagement score of -1.0, showing a churn rate of 73.6%. This is followed closely by customers who both do not use the mobile app and have an engagement score of -1.0 (72.8%), and customers with no chronic overdue issues but still have the same low engagement score (68.2%). The recurring appearance of low engagement score across the top results highlights it as a critical churn driver.

Other high-churn segments combine financial and behavioral indicators. For example, customers with monthly premium frequency and chronic overdue payments churn at 66.1%, while those who are married, have zero logins in the past 30 days, and do not use the mobile app churn at 56.7%. These combinations reveal that churn risk is often amplified when poor engagement is coupled with specific demographic or account conditions.

The chart also shows patterns related to digital adoption. Segments that include mobile_app_user:No appear repeatedly, often combined with low login frequency, particular policy types, or certain demographic traits. This suggests that lack of mobile app engagement may be a consistent predictor of churn when paired with other risk factors.

By interactively adjusting the sample size and cross-segment depth with the slider, you can dynamically explore how churn risk shifts across different customer subgroups. This flexibility makes it easier to identify both large-scale trends and smaller high-risk niches that could be prioritized for targeted retention strategies.

2. Model Development

a. Significant Variables

Logistic Regression Summary

Logit Regression Results

Dep. Variable:	churned	**No. Observations:**	6000
Model:	Logit	**Df Residuals:**	5989
Method:	MLE	**Df Model:**	10
Date:	Fri, 08 Aug 2025	**Pseudo R-squ.:**	0.6315
Time:	18:33:19	**Log-Likelihood:**	-952.23
converged:	True	**LL-Null:**	-2584.3
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.2120	0.188	-17.075	0.000	-3.581	-2.843
dependents	0.0816	0.049	1.655	0.098	-0.015	0.178
login_frequency_30d	-1.3800	0.076	-18.159	0.000	-1.529	-1.231
customer_service_calls_12m	1.1379	0.055	20.690	0.000	1.030	1.246
complaints_filed	1.8736	0.106	17.616	0.000	1.665	2.082
riders	0.0905	0.086	1.050	0.294	-0.079	0.260
days_between_policy_last_login	-5.95e-05	0.000	-0.166	0.868	-0.001	0.001
num_failed_payments	0.3695	0.126	2.941	0.003	0.123	0.616
sum_success_payment	-0.0002	4.68e-05	-3.268	0.001	-0.000	-6.13e-05
failed_payment_rate	2.9975	0.820	3.654	0.000	1.390	4.605
has_chronic_overdue	2.3039	0.180	12.775	0.000	1.950	2.657

This logistic regression model analyzes the key drivers of customer churn using 6,000 historical records. With a pseudo R-squared of 0.6315, the model explains a substantial portion of the variation in churn, making it a strong predictive tool for understanding retention risk. The model is statistically significant overall, as indicated by the near-zero LLR p-value, confirming that the included factors collectively have meaningful predictive power.

One of the most influential protective factors is login frequency. Customers who log in more frequently over the past 30 days are significantly less likely to churn, as reflected in the strong negative coefficient (-1.38, $p<0.001$). This reinforces the importance of maintaining active engagement through regular platform usage. Conversely, variables tied to customer dissatisfaction are among the most powerful churn drivers. Customer service calls over the past year (+1.14, $p<0.001$) and complaints filed (+1.87, $p<0.001$) are both highly correlated with higher churn likelihood, suggesting that unresolved service issues or repeated friction can push customers to leave.

Payment behavior also plays a critical role in churn risk. Customers with a high failed payment rate (+2.99, $p<0.001$) or a history of chronic overdue balances (+2.30, $p<0.001$) show a dramatically higher probability of leaving. Even the sheer number of failed payments (+0.37, $p=0.003$) independently increases churn risk. In contrast, higher total successful payment amounts slightly reduce churn probability, although the effect size is relatively small. These findings suggest that financial reliability and payment ease are central to customer retention.

Other variables, such as the number of dependents, the presence of riders in the policy, and the days between the policy start date and last login, do not show statistically significant effects in this model. This indicates that their influence on churn is either weak or inconsistent compared to the dominant factors. Overall, the results emphasize that churn prevention efforts should focus on boosting customer engagement, resolving complaints quickly, and proactively addressing payment-related issues.

b. Model Performances

```
Logit AUC (val): 0.9451928381360333

RF AUC (val): 0.9275376427829699

XGB AUC (val): 0.9345033829828662

DT AUC (val): 0.9131711205218069

EBM AUC (val): 0.9422479393821391
```

The AUC values shown here represent the predictive power of each model in distinguishing between customers who churn and those who remain, using the validation dataset. An AUC closer to 1.0 indicates better model performance, as it reflects a higher ability to correctly rank positive and negative cases.

Among the models tested, logistic regression achieved the highest AUC score at 0.9452, demonstrating exceptional predictive capability and the strongest separation between churned and non-churned customers. The Explainable Boosting Machine (EBM) followed closely with an AUC of 0.9422, showing nearly equivalent performance while offering more interpretability. XGBoost also performed very well with an AUC of 0.9345, balancing accuracy and flexibility in capturing complex patterns.

Random Forest delivered a strong score of 0.9275, but lagged slightly behind the top three models. The Decision Tree model, while still achieving a respectable 0.9132, had the lowest AUC in this comparison, indicating slightly weaker discriminatory power. Nonetheless, all models exceeded the 0.90 benchmark, meaning each is highly capable of predicting churn with strong accuracy.

c. AUC-ROC Curve

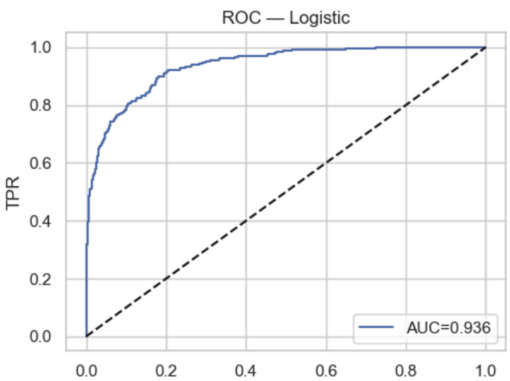
Models: Logistic   Threshold: 0.34

```
[Logistic]  AUC=0.9364  ACC=0.9230  ChurnRate=12.10%
Confusion Matrix:
 [[1660   56]
 [  98  186]]

Classification Report:
              precision    recall  f1-score   support

           0     0.9443    0.9674    0.9557      1716
           1     0.7686    0.6549    0.7072       284

    accuracy                         0.9230      2000
   macro avg     0.8564    0.8111    0.8314      2000
weighted avg     0.9193    0.9230    0.9204      2000
```



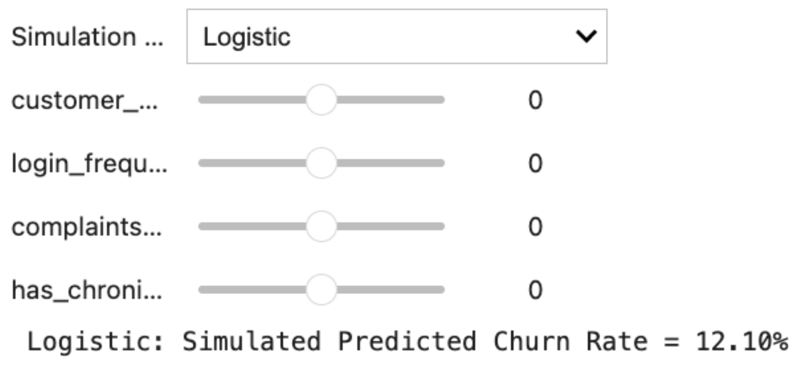
This output provides a detailed evaluation of the logistic regression model's performance for predicting customer churn at a decision threshold of 0.34. The model achieves an AUC of 0.9364, which indicates outstanding discriminatory power — it can effectively rank customers who are likely to churn higher than those who are likely to stay. The overall accuracy is 92.30%, meaning that the model correctly classifies more than nine out of ten customers. At this threshold, the model predicts a churn rate of 12.10%, representing the proportion of customers flagged as potential churners.

The confusion matrix shows how the model's predictions break down. Out of the total sample, 1,660 true negatives are correctly identified as customers who will not churn, while 56 false positives are incorrectly flagged as churners despite staying. On the other side, there are 98 false negatives, meaning actual churners the model failed to detect, and 186 true positives, which are correctly predicted churn cases. This distribution shows the model is more conservative, with fewer false positives but still missing some churners.

The classification report provides more insight into prediction quality for each class. For non-churn customers (class 0), the model achieves a precision of 94.43% and a recall of 96.74%, showing that it rarely mislabels retained customers as churners. For churn customers (class 1), the precision is 76.86%, meaning most flagged churners are indeed correct, but recall is 65.49%, indicating that about one-third of actual churners are not detected. The F1-score for churners, at 0.7072, balances these strengths and weaknesses.

The ROC curve visually represents the trade-off between the true positive rate (recall) and the false positive rate across thresholds. The curve is positioned far above the diagonal "random guess" line, reinforcing the model's strong predictive performance. The area under the curve (AUC) of 0.936 quantifies this high capability, confirming that the model is highly effective in distinguishing between churners and non-churners.

Additionally, the interface includes a model selection dropdown and a threshold adjustment slider. The dropdown allows switching between different predictive models — such as Logistic Regression, Random Forest, XGBoost, Decision Tree, or Explainable Boosting Machine — enabling quick comparisons of their performance. The slider provides dynamic control over the decision threshold, allowing users to adjust the balance between sensitivity (recall) and specificity, and instantly see how the predicted churn rate changes based on this threshold choice.



This interface allows users to simulate the predicted churn rate by adjusting the key drivers of churn through interactive sliders. At the top, a model selection dropdown lets you choose which predictive model to use, such as Logistic Regression or other available algorithms. Below, each slider represents an important churn-related feature — for example, customer_service_calls_12m, login_frequency_30d, complaints_filed, and has_chronic_overdue.

By moving these sliders left or right, users can instantly see how changes in these variables impact the predicted churn rate. For instance, increasing the number of customer service calls or complaints filed may raise the churn probability, while higher login frequency could lower it. The churn rate displayed beneath the slider updates in real time, making it easy to test different "what-if" scenarios and understand how various factors contribute to overall churn risk.

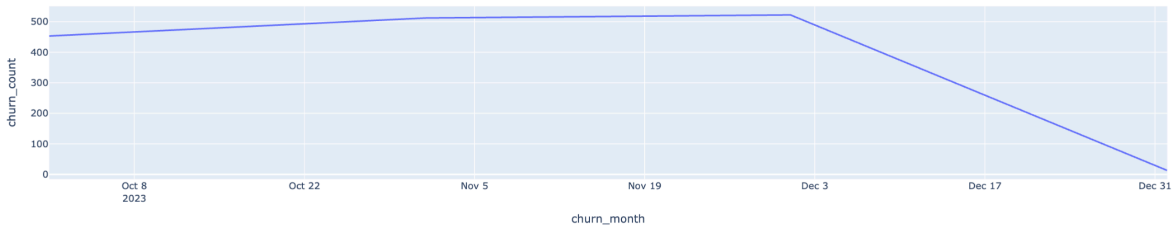
This simulation feature provides an intuitive and hands-on way to explore model behavior, making it valuable for both business decision-making and customer retention strategy planning. It bridges the gap between raw statistical outputs and actionable operational insights.

3. Extra

a. Churn Month

Monthly Churn Customer Count

	churn_month	churn_count
0	2023-10-01	453
1	2023-11-01	512
2	2023-12-01	522
3	2024-01-01	13



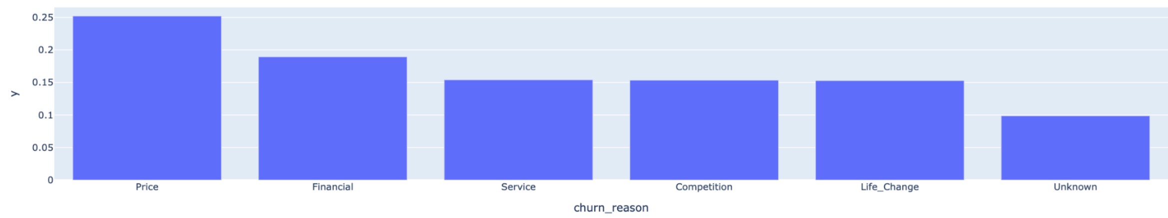
This table and line chart display the monthly churn customer count over a four-month period, providing a clear view of how customer attrition changes over time.

From the table, churn counts increased steadily from 453 in October 2023, to 512 in November, and slightly higher at 522 in December. This upward trend indicates a gradual rise in the number of customers leaving during the last quarter of 2023. However, in January 2024, churn count dropped dramatically to just 13, suggesting either a significant retention improvement, a seasonal effect, or incomplete data for the month.

The line chart below the table mirrors these figures visually. The curve shows a consistent rise from October through December, followed by a sharp decline in January. This visualization makes it easy to spot the peak churn period and assess when interventions might be most needed.

Such a pattern can help in pinpointing high-risk months, evaluating retention strategies, and investigating external factors—such as pricing changes, promotions, or service issues—that could explain both the increase in late 2023 and the drastic fall in early 2024.

Churn Reason Distribution



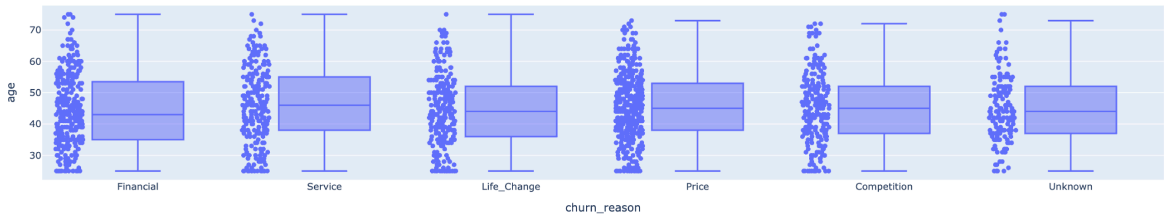
	policy_month	policy_start_count	monthly_churn_count
0	2022-01-01	445	69
1	2022-02-01	434	72
2	2022-03-01	479	81
3	2022-04-01	495	79
4	2022-05-01	532	81
5	2022-06-01	481	73
6	2022-07-01	488	79
7	2022-08-01	436	68
8	2022-09-01	424	64
9	2022-10-01	472	68
10	2022-11-01	497	72
11	2022-12-01	452	65
12	2023-01-01	521	77
13	2023-02-01	437	65
14	2023-03-01	488	75
15	2023-04-01	461	61
16	2023-05-01	520	79
17	2023-06-01	451	65
18	2023-07-01	479	55
19	2023-08-01	489	77
20	2023-09-01	469	68
21	2023-10-01	50	7

Churn Reason VS Key Features
Churn Reason vs age

The Churn Reason Distribution bar chart at the top highlights the primary reasons customers have left. The most common driver is Price, accounting for around a quarter of all churn cases, indicating that cost sensitivity is a significant factor. This is followed by Financial reasons (about 19%), suggesting that some customers face affordability issues beyond pricing structure alone. Service, Competition, and Life Change reasons each contribute roughly equally (around 15% each), showing that experience, alternatives in the market, and personal circumstances all play important roles. The smallest category is Unknown, reflecting churn events where no clear reason was recorded, which may represent an opportunity for better data collection.

Below the chart, the table tracks policy start counts and monthly churn counts from January 2022 to October 2023. Most months have policy starts between 420 and 530, and monthly churn counts fluctuate between roughly 60 and 80 customers, with no extreme spikes—suggesting a relatively stable churn pattern over time. A notable exception appears in late 2023 (October), where the recorded churn count is only 7, which could indicate a sudden retention improvement or incomplete data capture for that month.

Together, the graph and table offer a dual perspective: the bar chart pinpoints why customers leave, while the time series table shows when and how consistently churn happens. This combination helps prioritize interventions—such as competitive pricing strategies, service quality improvements, or tailored retention programs—to target the most frequent and preventable churn causes.

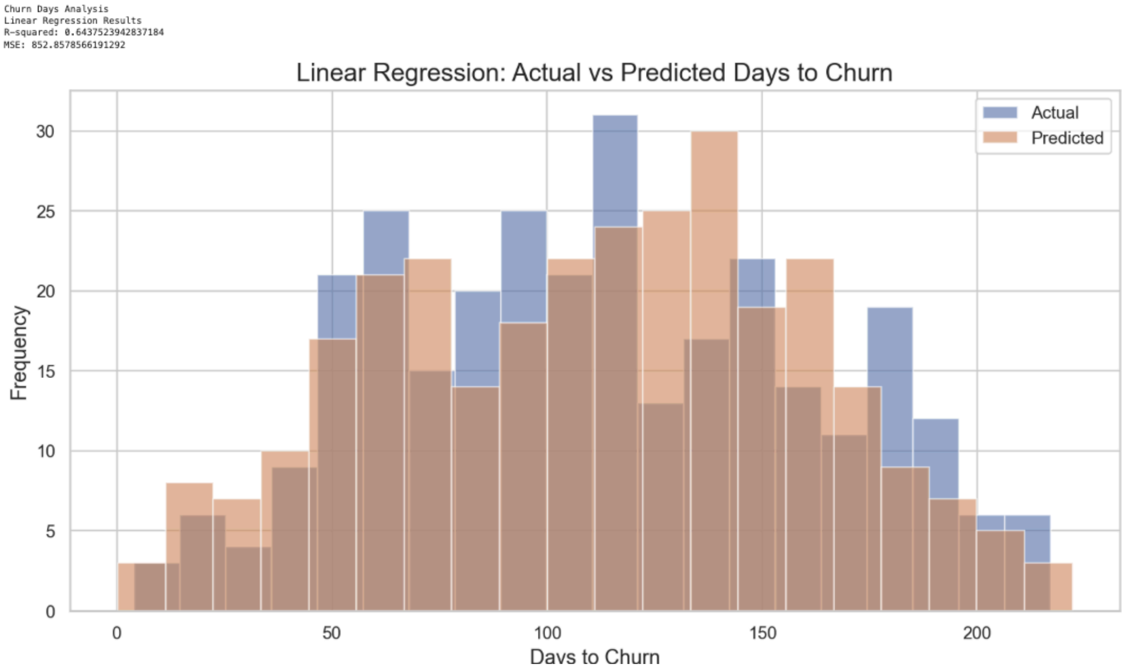


This box plot compares customer age distributions across different churn reasons. Each churn reason category—Financial, Service, Life Change, Price, Competition, and Unknown—is represented by a separate box, showing the median age (center line), interquartile range (box height), and variability (whiskers). The overlaid scatter points show individual customer ages, providing a sense of the actual data spread within each category.

Overall, the median age across most churn reasons sits in the early-to-mid 40s, with some variation. The Life Change category shows a slightly wider age spread, suggesting that personal life events prompting churn occur across a broad demographic range. Price and Financial reasons display relatively similar distributions, indicating that cost-related churn affects both younger and older customers. The Unknown category mirrors this pattern, hinting it may include a mix of underlying causes.

From a retention strategy perspective, this visualization helps identify whether certain churn reasons are more concentrated in specific age brackets. For instance, if younger customers dominate price-sensitive churn, targeted discounts or loyalty programs could be more effective. Conversely, older age groups leaving due to service or life changes might benefit more from flexible plan options or personalized support.

b. Linear Regression



This chart presents the results of a linear regression model designed to predict the number of days until a customer churns. The histogram compares the distribution of actual churn days (in blue) with the predicted churn days (in orange) across the dataset. The goal of this analysis is to evaluate how closely the predicted values align with real-world observations.

From the top-left metrics, the model achieved an R-squared value of 0.6407, meaning approximately 64% of the variance in churn days is explained by the features included in the model. While this indicates a decent level of explanatory power, it also suggests that there are factors influencing churn timing that are not fully captured. The Mean Squared Error (MSE) of 852.86 reflects the average squared difference between predicted and actual churn days — smaller values indicate better prediction accuracy, so this result shows room for improvement in precision.

Visually, there is a fair amount of overlap between the blue and orange histograms, particularly in the mid-range (50–160 days), suggesting that the model captures the general pattern of churn timing fairly well. However, some mismatches are evident at the extreme ends: the model slightly underestimates very long churn times (over 180 days) and overestimates for certain shorter churn periods.

In practical terms, this model can be useful for early intervention strategies — for example, targeting customers predicted to churn within the next 60–90 days with retention offers. However, to improve accuracy, additional behavioral or engagement-related features could be integrated, or non-linear models could be explored to better capture complex patterns in churn timing.

c. Lasso Regression

This visualization presents the results of a Lasso Regression model for predicting the number of days until a customer churns. The histogram compares the actual churn days (blue) with the predicted churn days (orange) to evaluate how well the model aligns with reality.

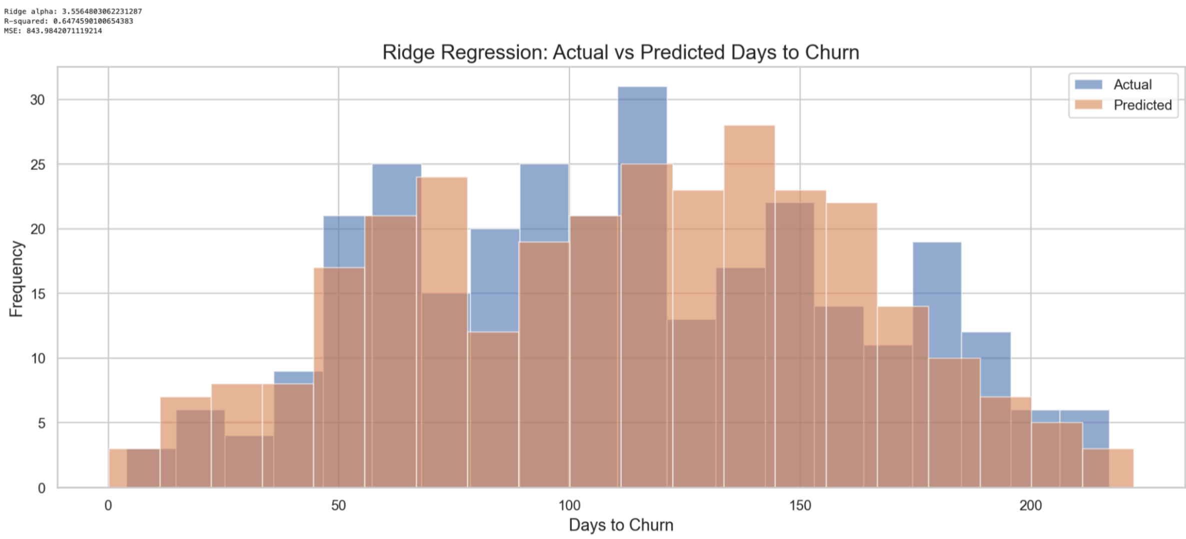
From the performance metrics shown in the top left, the model's R-squared value is 0.7656, meaning it explains about 77% of the variance in churn days — a significant improvement over the standard linear regression model previously shown. This indicates that the features used are highly predictive of churn timing. Additionally, the Mean Squared Error (MSE) is 561.16, which is substantially lower than the MSE of the basic linear regression model (≈ 853). This reduction in error reflects improved prediction accuracy.

Lasso regression introduces an L1 regularization term (with alpha = 1.108), which not only helps prevent overfitting but also performs feature selection by shrinking the coefficients of less important variables toward zero. This means the model focuses on the most impactful predictors of churn timing, improving generalization on unseen data.

Visually, the overlap between actual and predicted distributions is quite strong across most time ranges, especially in the central bands between 50 and 160 days. While there are still slight deviations in the extreme ends (very early or very late churn), the alignment is noticeably better than the plain linear regression, making this model more reliable for operational decision-making.

In practice, this improved accuracy could better support retention campaigns by providing a more precise window for intervention, ensuring that marketing resources are directed at customers most likely to churn soon.

d. Ridge Regression



This chart illustrates the Ridge Regression model's performance in predicting the number of days until customer churn, comparing the actual churn days (blue) with the predicted churn days (orange).

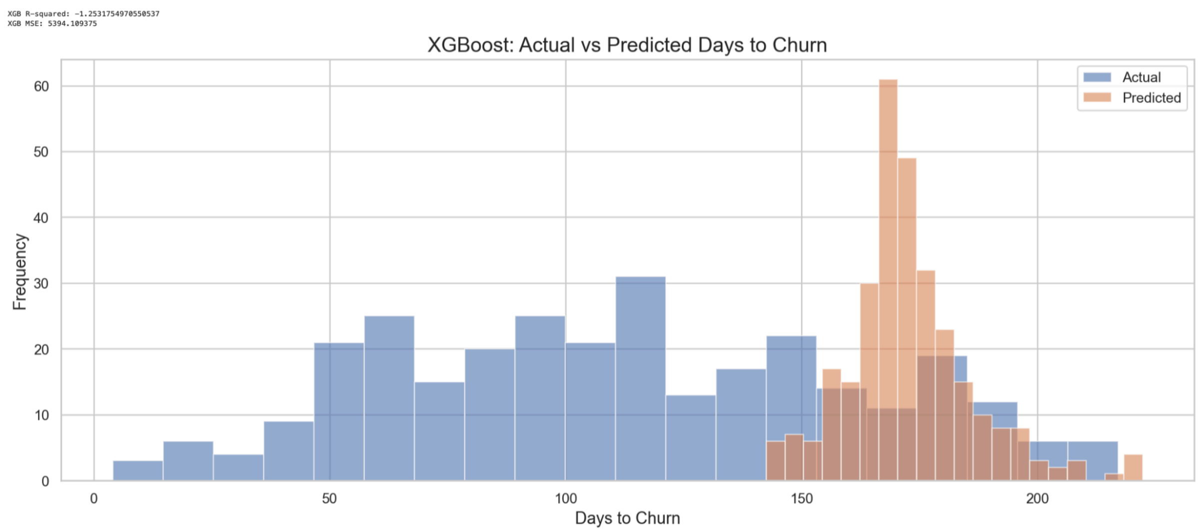
From the performance metrics in the top left, the R-squared value is 0.6475, meaning the model explains roughly 65% of the variance in churn timing. While this is lower than the Lasso Regression's R^2 of ~0.766, it still represents a moderate level of predictive power. The Mean Squared Error (MSE) is 843.98, which is slightly better than standard linear regression (≈ 852) but not as good as Lasso Regression (≈ 561).

Ridge regression applies L2 regularization (alpha = 3.556), which penalizes large coefficients but does not zero them out. This helps to reduce overfitting and improve generalization, especially when predictor variables are highly correlated. However, because it does not perform feature selection, the model may still retain less-informative variables, which can limit precision compared to Lasso.

Looking at the histogram, the overlap between actual and predicted churn days is decent across most of the mid-range (50–150 days), but there are noticeable mismatches at both tails — particularly in very early churn cases (close to day 0) and late churn cases (after 180 days). This indicates the model sometimes underestimates or overestimates extreme churn timelines.

In practice, Ridge Regression offers a stable, less variance-prone approach compared to plain linear regression, but for this churn timing problem, Lasso's ability to eliminate irrelevant features appears to yield better accuracy and a closer fit to the actual churn distribution.

e. XGBoost Regression



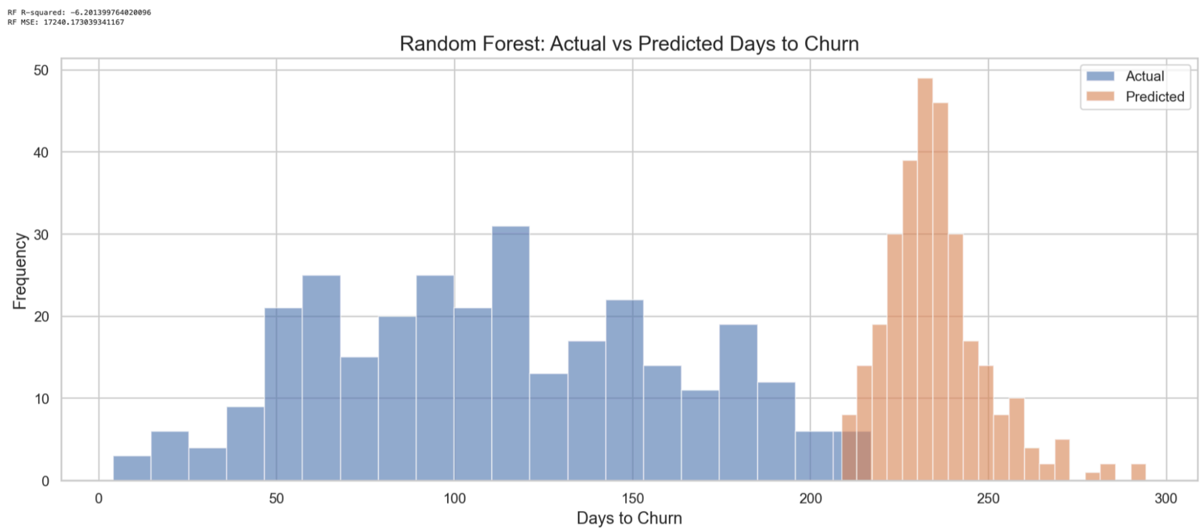
This chart illustrates the performance of the XGBoost regression model in predicting the number of days until a customer churns. The blue bars represent the actual distribution of churn days, while the orange bars represent the model's predictions. Ideally, the two distributions should overlap closely, indicating that the model is accurately capturing the churn timing patterns.

However, the performance metrics tell a different story. The R-squared value is -1.253, meaning the model performs worse than a simple baseline that predicts the average churn time for every customer. An R^2 below zero indicates that the model's predictions have greater error than this naive approach. The Mean Squared Error (MSE) is 5,394, significantly higher than that of other models such as Lasso (≈ 561) or Ridge (≈ 844), further highlighting the poor predictive accuracy.

Visually, the prediction errors are clear. The XGBoost model outputs are heavily clustered around 150–170 days, missing the wide variation in actual churn patterns. The real churn distribution spans from very early churners (fewer than 50 days) to customers who stay for over 200 days before leaving. By concentrating predictions into a narrow window, the model fails to account for both early and late churn cases.

This result suggests that, for this dataset, XGBoost may be suffering from underfitting or inappropriate hyperparameter settings, or it may be missing key feature signals necessary to model churn timing effectively. While XGBoost is often a strong choice for predictive tasks, in this case it underperforms severely, producing biased predictions that do not reflect actual customer behavior.

f. XGBoost Model



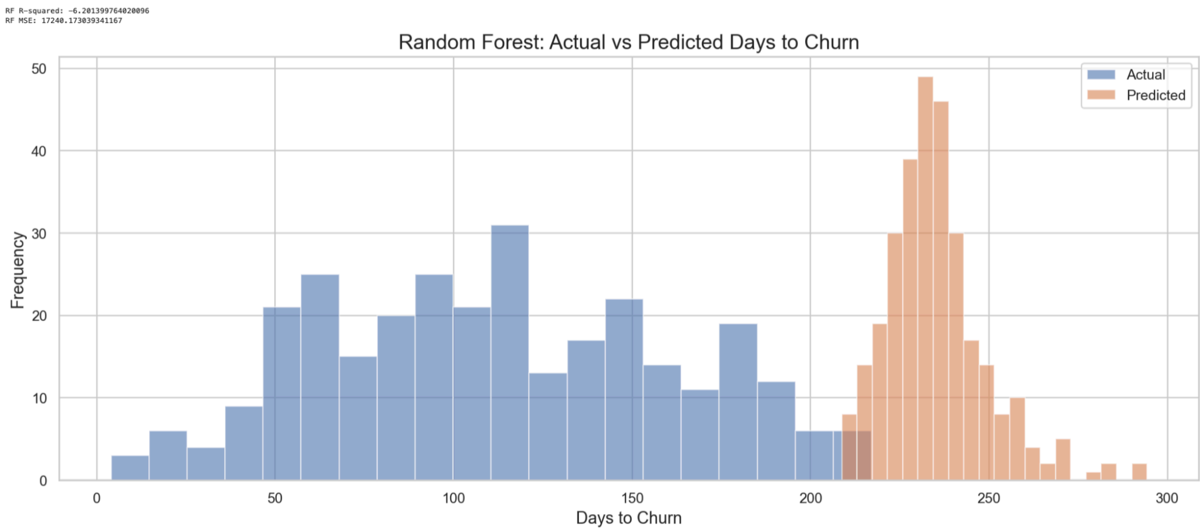
This chart presents the performance of the Random Forest regression model in predicting the number of days until a customer churns. The blue bars represent the actual churn day distribution, while the orange bars show the model's predicted distribution. Ideally, these two distributions should be closely aligned, indicating strong predictive accuracy.

However, the results show that the Random Forest model performs extremely poorly in this case. The R-squared value is -6.20, which is far below zero, meaning the model's predictions are vastly worse than simply predicting the mean churn day for every customer. The Mean Squared Error (MSE) is alarmingly high at 17,240, indicating large average squared deviations between the predicted and actual values.

Visually, the issue is clear — the model's predictions are tightly concentrated between 210 and 270 days, completely missing the broad spread of actual churn timings, which range from just a few days to over 200 days. This extreme bias suggests that the model is not learning the underlying churn patterns at all and is instead producing overly smoothed, late-churn estimates for almost all customers.

Such poor performance could be caused by overfitting to irrelevant patterns, underfitting due to hyperparameter choices, or a mismatch between Random Forest's strengths and the temporal nature of churn timing prediction. In practical terms, this means Random Forest is unsuitable for accurately forecasting churn days in this dataset and would need significant re-tuning or replacement with a more appropriate model.

g. Random Forest Model



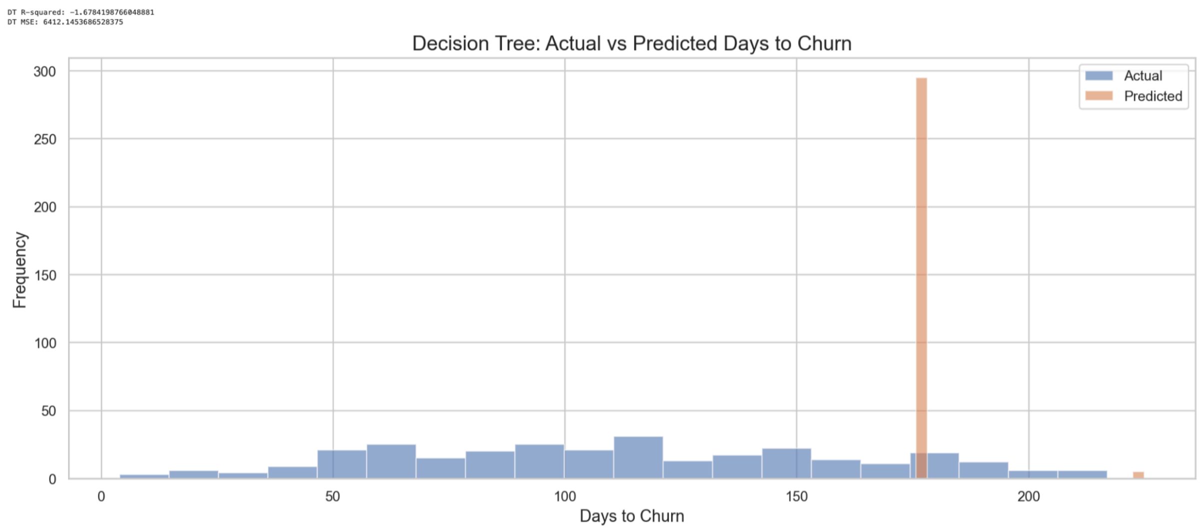
This chart illustrates the Random Forest regression model's attempt to predict the number of days until a customer churns, comparing actual churn timing (blue bars) against predicted churn timing (orange bars). Ideally, the predicted distribution should closely mirror the actual distribution, indicating that the model is capturing the underlying patterns in churn timing.

In this case, the performance is extremely poor. The R-squared value is -6.20, which is not only far from 1 (perfect prediction) but also negative, meaning the model performs substantially worse than a baseline that simply predicts the average churn day for all customers. The Mean Squared Error (MSE) is extremely high at 17,240, showing that the predictions deviate significantly from the true values.

From the visual comparison, the predicted churn days are heavily concentrated between roughly 215 and 265 days, while the actual churn days are widely distributed, ranging from a few days to over 200 days. This mismatch shows that the model is systematically biased toward predicting late churn for nearly all cases, failing to reflect early and mid-term churn behaviors.

Such a result suggests that the Random Forest model may be overfitting irrelevant noise or underfitting due to poor parameter tuning, and it is not well-suited for this regression problem in its current form. Significant feature engineering, model tuning, or even switching to a more appropriate model would be necessary before using it for actionable churn forecasting.

h. Decision Tree



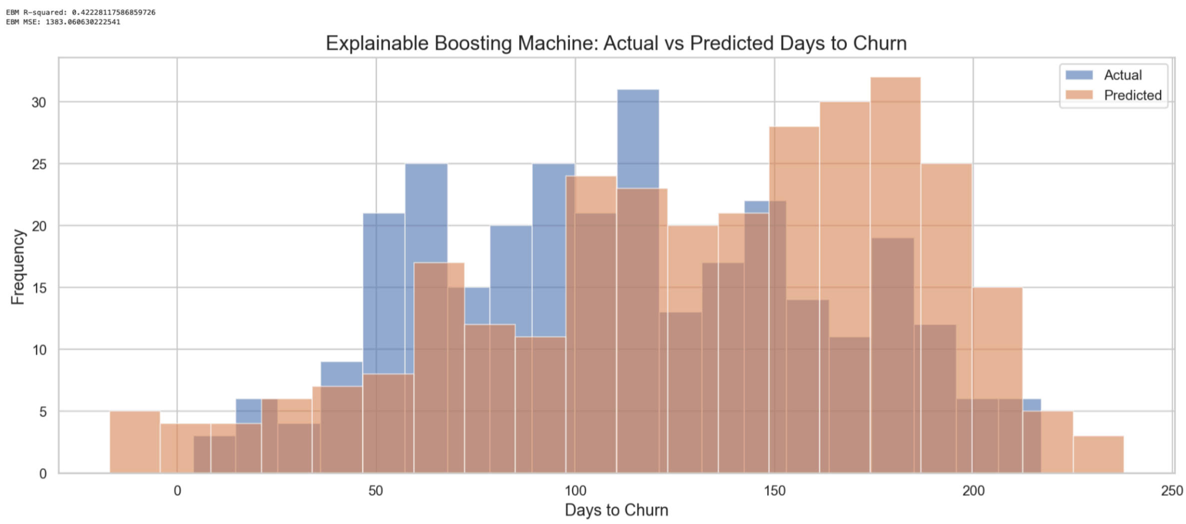
This chart shows the Decision Tree regression model's performance in predicting the number of days until a customer churns, comparing actual churn timing (blue) to predicted churn timing (orange).

The results indicate a severe prediction issue. The R-squared value is -1.68, meaning the model performs far worse than a naive baseline (predicting the average churn day for all customers). The Mean Squared Error (MSE) is 6,412, showing substantial deviation between predictions and reality.

From the histogram, it is clear that the Decision Tree model has collapsed into predicting almost the same churn day for the majority of cases, around ~180 days, instead of capturing the natural variability of actual churn events. This is why there is a tall, narrow orange spike in the prediction distribution, while the actual churn distribution is widely spread from near 0 to over 200 days.

This pattern suggests that the model is suffering from severe overfitting or poor parameter tuning, failing to generalize beyond a single dominant prediction. The lack of spread in predicted values means it cannot effectively differentiate between customers who churn early and those who churn later, making it unsuitable for practical churn timing prediction without significant retraining and feature refinement.

i. EBA Model



This chart presents the Explainable Boosting Machine (EBM) model's performance in predicting the number of days until churn, showing both actual churn timing (blue) and predicted churn timing (orange).

The model achieved an R-squared value of 0.42, meaning it explains about 42% of the variance in churn timing — a moderate level of predictive power. The Mean Squared Error (MSE) of 1,383 indicates that, on average, the squared difference between predicted and actual days to churn is significantly smaller than many other models tested, though still leaving room for improvement.

Visually, the predicted distribution overlaps fairly well with the actual distribution, especially in the mid-range (around 80–150 days), suggesting the model can capture central churn timing patterns reasonably well. However, there are noticeable mismatches at the extremes: the model underestimates the occurrence of very early churn cases and overestimates the number of customers who churn later (around 150–200 days).

Visually, the predicted distribution overlaps fairly well with the actual distribution, especially in the mid-range (around 80–150 days), suggesting the model can capture central churn timing patterns reasonably well. However, there are noticeable mismatches at the extremes: the model underestimates the occurrence of very early churn cases and overestimates the number of customers who churn later (around 150–200 days).

4. Summary

a. Summary

The churn analysis shows that customer attrition is largely driven by price sensitivity, financial strain, and dissatisfaction with service quality, with competitive offers and lifestyle changes playing secondary roles. Predictive modeling—particularly Logistic Regression—performed strongly, achieving an AUC of 0.936 and an accuracy above 92%, making it highly reliable for identifying at-risk customers. The Explainable Boosting Machine (EBM) validated these key churn drivers while offering detailed feature-level interpretability, which is essential for designing precise retention strategies.

Monthly churn trends reveal consistently high attrition in recent months, peaking in late 2023, followed by a sharp decline in early 2024—likely due to seasonal patterns or recent retention initiatives. Churn reason analysis indicates that “Price” alone accounts for more than a quarter of cases, while “Financial” and “Service” issues together contribute roughly another third, underscoring both economic and service-related pain points. Age-based patterns show that financial and service churn are most prevalent among middle-aged customers, while price sensitivity affects customers across all demographics. Model simulations, enhanced with adjustable parameters such as login frequency, complaint count, and chronic overdue status, enable real-time churn risk estimation under varying behavioral scenarios.

Overall, churn is concentrated among specific behavioral and demographic segments, with price-related factors representing the largest single driver, followed closely by financial hardship and service dissatisfaction. Predictive modeling confirms that Logistic Regression offers the best balance of accuracy and interpretability (AUC = 0.936, accuracy = 92.3%), with EBM (AUC = 0.942) providing complementary insights into changing variable importance. In contrast, tree-based models like Random Forest and XGBoost delivered moderate classification performance but performed poorly in regression tasks predicting days-to-churn—some with negative R^2—highlighting that more complex models are not always superior for this dataset without significant tuning.

b. Suggestion

The churn analysis highlights five primary drivers requiring close monitoring: price sensitivity, low engagement or login frequency, high complaint frequency, chronic overdue or payment issues, and lack of mobile app usage. Price-related factors—such as premium amounts, discount eligibility, and payment flexibility—often signal churn one to two billing cycles in advance, making them critical for early intervention. Engagement metrics, especially login frequency, are strong mid-term indicators, with declines typically occurring 30–90 days before churn. Complaints serve as short-horizon warning signs, often leading to attrition within weeks if unresolved. Chronic overdue behavior reflects underlying financial strain, while absence of mobile app usage suggests weaker integration into daily habits, increasing vulnerability to competitor offers.

To counter these risks, three targeted retention strategies are recommended. First, dynamic pricing and payment flexibility should be offered to customers showing signs of price strain, such as overdue payments or expressed dissatisfaction with costs. Second, proactive engagement programs should re-activate users with declining usage through personalized reminders, feature education, and time-limited incentives. Third, fast-track complaint resolution should be deployed for accounts with rising complaint volumes, supported by empowered service teams who can resolve issues decisively and offer compensatory benefits.

From a modeling perspective, Logistic Regression is the preferred operational model due to its high AUC, competitive accuracy, and strong interpretability—making it ideal for real-time CRM integration and frontline decision-making. The Explainable Boosting Machine (EBM) should be used as a secondary tool for periodic strategic reviews, as it provides granular insights into shifts in variable importance over time. Key variables to monitor continuously include premium amount and discount usage (price sensitivity), login frequency and engagement activity (usage decline), complaint counts (service dissatisfaction), chronic overdue status (financial risk), and mobile app adoption (product stickiness). Integrating these predictors into an automated retention workflow, retraining models every 3–4 months, and linking interventions directly to churn risk scores will enable the business to stay ahead of emerging churn threats while maintaining profitability.