GreenwallTech Solutions
Churn Analysis

Joshua Thompson
Date: 08/17/2025

Table of Contents

1. Executive Summary

Business Context

GreenwallTech Solutions, a SaaS provider, has seen a decline in annual customer retention, threatening recurring revenue. Leadership needs clear, data-driven insights to understand why customers churn and how to prevent it.

The company's integrated "gold" layer dataset in Microsoft Fabric contains demographics, subscription history, engagement metrics, support records, and churn labels. Leveraging this data, a predictive churn model can:

- Identify the strongest churn drivers.
- Flag high-risk customers for proactive retention.
- Guide targeted business interventions to improve loyalty and lifetime value.

A successful solution will not only help recover retention rates but also optimize resource allocation and strengthen long-term customer relationships.

Model Development

- Data: 5,000 customers from the company's "gold" layer dataset (demographics, subscriptions, engagement, support, churn labels).
- Preprocessing: Outlier clipping, scaling, one-hot encoding for categorical features, and variance checks ensured clean inputs.
- Feature Engineering: Introduced spend_per_tenure (monthly spend ÷ tenure) to better capture short-term customer value, avoiding inflated lifetime value estimates.
- Multicollinearity: All features showed low VIF (< 2.2), confirming no collinearity concerns.

Model Results

Chosen Model: Logistic Regression selected over Decision Tree and SVM due to interpretability, stable performance, and linear feature relationships.

Performance:

- Churn recall: 0.81 (captures ~81% of churners)
- Churn precision: 0.46 (about half of flagged churners are true churners)
- ROC AUC: 0.62 (moderate discrimination)

Threshold Tuning:

- Optimal F1-score achieved at threshold ~0.34.
- At 0.44 threshold: Churn recall improved to 81% but precision dropped to 46%. This reflects a trade-off between catching more churners vs. limiting false alarms.

Key Churn Drivers

Higher Churn Risk	Lower Churn Risk
• Basic subscription (+30% odds of churn). • Longer time since last login (+29%). • Higher support ticket volume (+14%). • Customers in East & North regions (+3–5%).	• High monthly spend (−14% churn odds). • Longer tenure (−21%). • Standard subscription type (−20%). • Customers in West region (−7%).

Business Insights

- Churn is most strongly linked to engagement, subscription type, and support experience.
- High-value, long-tenure customers are naturally “stickier,” while Basic-plan and disengaged customers represent the most urgent churn risks.
- The model ranks customers meaningfully: the top 10% of predicted churn risk are 1.8× more likely to churn than average, enabling efficient resource allocation.

Overall Takeaway: The churn model delivers actionable segmentation, balancing interpretability and predictive accuracy. It provides a foundation for proactive retention campaigns, particularly targeting disengaged Basic-plan customers with frequent support issues, while allowing leadership to prioritize intervention resources where they matter most.

Limitations & Potential Improvements

- **Moderate Predictive Power:** ROC AUC of 0.62 indicates the model has only moderate discriminative ability. While it captures a majority of churners (high recall), precision is limited, meaning many flagged customers may not actually churn.
- **Class Imbalance Sensitivity:** Although class weights were balanced, churn remains a relatively rare event. More advanced methods, with gridsearch hyperparameter tuning could improve robustness.
- **Optimization of Business Outcomes:** Current thresholds maximize F1-score, but future iterations could optimize directly for financial impact balancing retention campaign costs against expected revenue saved.

2. Exploratory Analysis

2.1 Categorical Features

	Data Type	Missing (%)	# Unique	Most Frequent	Freq (%)
region	object	0.0	4	South	25.78
subscription_type	object	0.0	3	Standard	40.72
churned	category	0.0	2	0	62.56

Key Insights:

- Region Distribution: No single region dominates overwhelmingly, but “South” is the largest segment; may be worth checking if churn varies by region.
- Subscription Concentration: Standard plan is the largest group, which could drive the overall churn pattern if churn is uneven across plans.
- Churn Distribution: Majority of customers are retained (~63%), but churn is still high enough to justify targeted interventions.

No Missing data. No need for imputation methods

2.2 Numerical Features

	Data Type	Missing (%)	Mean	Median	Min	25th %ile	75th %ile	Max	Skewness	Outlier (%)
age	int64	0.00	43.49	44.00	18.0	31.00	56.00	69.00	0.01	0.00
tenure_months	int64	0.00	30.20	30.00	1.0	15.00	45.00	60.00	0.04	0.00
last_login_days_ago	int64	0.00	49.62	50.00	0.0	25.00	75.00	99.00	0.00	0.00
num_support_tickets	int64	0.00	1.52	1.00	0.0	1.00	2.00	7.00	0.79	6.96
monthly_spend	float64	0.00	53.64	49.91	10.0	28.36	73.91	119.99	0.45	0.00
engagement_score	float64	0.00	90.71	100.00	0.0	95.37	100.00	100.00	-2.32	21.18

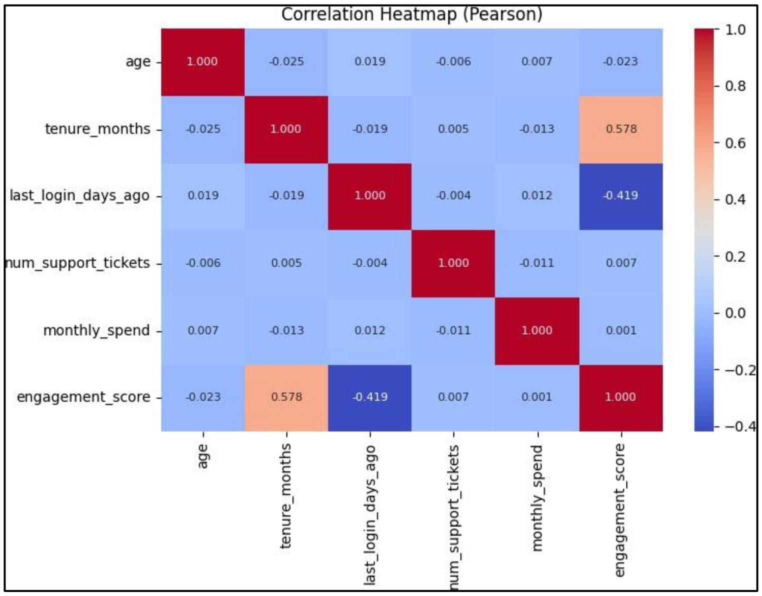
Key Insights:

- Engagement Score: Strong negative skew; most customers have very high engagement, but a notable minority have very low scores (~21% outliers).
- Monthly Spend: Mild positive skew; most customers spend moderately, with a small portion spending significantly more.
- Support Tickets: Mild positive skew; the majority raise few tickets, while ~7% are frequent support users.
- Tenure: Balanced distribution centered around ~30 months, with both short- and long-tenure customers represented.
- Age: Evenly distributed around the mean of 43 years; no extreme skew.
- Last Login: Symmetric distribution, median of 50 days ago.

No Missing data. No need for imputation methods

2.3 Correlation Matrix

Next, I examined feature correlations to identify potential multicollinearity. This helps ensure the model isn't biased by redundant or highly correlated variables.



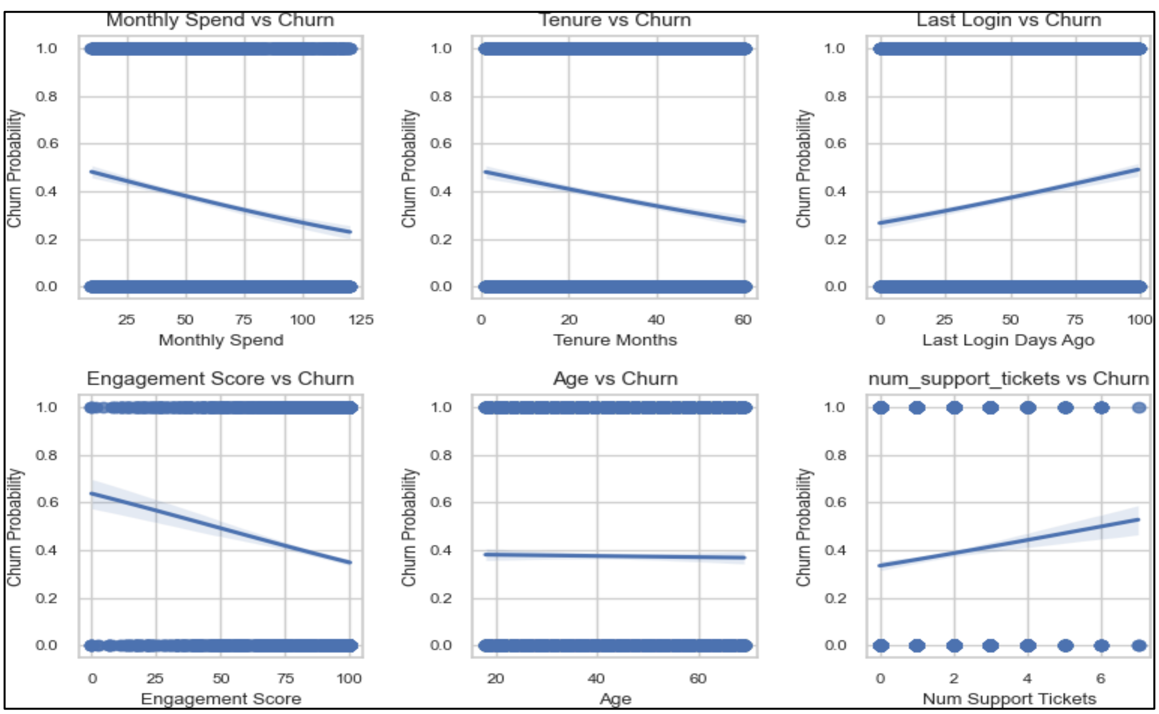
Key Insights

Strongest Relationships

- Tenure & Engagement Score: 0.578 are Customers with longer tenure tend to have higher engagement. This is the strongest positive correlation in the dataset.
- Last Login Days Ago & Engagement Score: -0.419 are Customers who logged in more recently tend to have higher engagement. This is the strongest negative correlation.

2.4 Feature Relationships with Churn Probability

Before proceeding to model development, I conducted a logistic regression to examine the propensity to churn across different feature values. This helped identify which variables exhibit a linear relationship with churn probability, providing early insight into feature behavior and informing the choice of logistic regression as a suitable baseline model.



Key insights

1. Visual Linearity Suggests Logistic Regression as a Baseline Several features including: monthly_spend, tenure_months, and engagement_score, exhibit a reasonably linear trend with respect to churn. This justifies starting with a linear based model like logistic regression as an initial baseline model, offering both interpretability and a solid foundation for performance comparison.

2. Lack of Strong Non-Linear Patterns No significant curvilinear or threshold patterns are apparent from the plots. This suggests that simpler models may perform competitively. Although I will test other models such as Random forest, Naive bayes, etc.. for sense check comparison.

---

3. Preprocessing & Transformation Pipeline

To prepare the dataset for churn modeling, I implemented a two-branch preprocessing pipeline: one for numeric features and one for categorical features. The combined pipeline ensures clean, scaled, and model-ready data.

1. Numeric Feature Pipeline
   - Low Variance Filter: Removes features with minimal variability (threshold = 0.01), which are unlikely to contribute meaningfully to model predictions.
   - Outlier Clipping: Caps extreme values at the 1st and 99th percentiles to reduce the influence of outliers and improve model robustness.
   - Standard Scaling: Standardizes all numeric features to zero mean and unit variance to ensure uniform feature scaling, especially important for distance-based or regularized models.

2. Categorical Feature Pipeline

   - One-Hot Encoding: Converts categorical variables into binary indicator variables. handle_unknown='ignore' ensures the pipeline is robust to unseen categories during inference.

3. Combined Transformation (ColumnTransformer)

Applies the numeric pipeline to all numeric columns and the categorical pipeline to categorical columns in parallel, producing a unified transformed dataset ready for machine learning models.

3.1 Split – Train (70%) / Validation (15%) / Test (15%)

Next, to ensure robust and unbiased evaluation, I split the dataset in two stages:

- Test Set (15%): A hold-out set was first separated to serve as an untouched benchmark for final model evaluation.

- Train/Validation Split (85%): The remaining data was split into training (70%) and validation (15%) sets. Since 15% of the full dataset represents approximately 17.6% of the remaining 85%, this was done using test_size=0.176.

This approach preserves class balance using stratified sampling, prevents data leakage, and ensures that model evaluation is based on data the model has never seen, mimicking real-world deployment performance.

---

4. Model Building

4.1 Stage 1: Baseline model

I begin with a baseline model to establish a performance benchmark. This helps validate the end to end pipeline, uncover basic patterns, and ensures that any improvements from more complex models are meaningful and justified.

=== Cross-Validation Results (Models in Columns) ===

	Logistic Regression
roc_auc	0.648 ± 0.026
pr_auc	0.517 ± 0.026
f1	0.526 ± 0.016
recall	0.587 ± 0.025
precision	0.476 ± 0.012

Baseline Model Evaluation (Logistic Regression)

- **Recall: 0.587 ± 0.025** -About 59% of actual churners are being identified, but it may still miss many churners.
- **Precision: 0.476 ± 0.012** -Less than half of the predicted churns are correct, so there's a high false positive rate. This could lead to unnecessary retention efforts if used in production as is.
- **ROC AUC: 0.648 ± 0.026** -The model distinguishes between churned and non-churned customers better than random (0.5), but there's still significant room for improvement.

4.2 Stage 2: Feature Engineering & Model Optimization

In this stage, I enhance the dataset through targeted feature engineering, followed by training multiple model types to compare performance. I then apply hyperparameter tuning to optimize the best-performing models, ensuring improved accuracy and generalizability over the baseline.

Feature Engineering

To better understand which customers are most valuable, I initially considered using a traditional lifetime value metric calculated as ltv = monthly_spend × tenure_months. However, I realized this approach could inflate values for long-tenured customers, since their LTV would naturally appear higher simply due to time rather than true spending intensity.

Instead, I created a new metric: spend_per_month = monthly_spend ÷ tenure_months. This captures how much a customer spends relative to their tenure, surfacing high-value customers who may not have been around long but are already spending heavily. By balancing spend and tenure, this feature provides a clearer view of short-term value contribution and allows more nuanced retention and outreach strategies.

Variance Inflation Factors - Check

After creating this new feature, I assessed potential multicollinearity by calculating Variance Inflation Factors (VIF). This step helps identify highly correlated features that could bias the model and ensures more reliable coefficient estimates.

	feature	VIF
0	const	64.598115
1	age	1.001247
2	tenure_months	1.754544
3	last_login_days_ago	1.365821
4	num_support_tickets	1.000288
5	monthly_spend	1.110562
6	engagement_score	2.189253
7	spend_per_tenure	1.515828

VIF Summary

- All feature VIF values are well below the common thresholds of concern (~5).
- The highest VIF is for engagement_score (2.19), which is still low and not problematic.
- Newly engineered feature spend_per_tenure (1.52) shows no signs of excessive collinearity with existing features.

Overall, no multicollinearity issues are present. Which means all features can be retained without needing adjustment.

4.3 Model Comparison

With feature engineering complete, the next step is to build a modeling pipeline where I will test the new feature, compare the performance of multiple models, and then make a final model selection based on results.

=== Cross-Validation Results (Models in Columns) ===

	Logistic Regression	Decision Tree	SVM (RBF Kernel)	Random Forest	XGBoost
roc_auc	0.646 ± 0.028	0.617 ± 0.028	0.642 ± 0.029	0.643 ± 0.031	0.622 ± 0.023
pr_auc	0.514 ± 0.027	0.464 ± 0.024	0.496 ± 0.027	0.498 ± 0.028	0.487 ± 0.031
f1	0.530 ± 0.022	0.528 ± 0.035	0.537 ± 0.036	0.503 ± 0.027	0.509 ± 0.018
recall	0.596 ± 0.034	0.608 ± 0.091	0.599 ± 0.047	0.506 ± 0.027	0.542 ± 0.010
precision	0.478 ± 0.017	0.475 ± 0.036	0.487 ± 0.030	0.501 ± 0.033	0.480 ± 0.026

Final Model Selection: Logistic Regression

I decided to move forward with Logistic Regression as the final model due to its strong interpretability, stable performance, and alignment with the structure of the data. While other models like SVM, Decision Tree and ensemble methods showed similar F1 and recall scores, Logistic Regression offers transparent coefficient outputs, making it easier to explain feature influence on churn to stakeholders. Given the relatively small number of features, lack of complex nonlinear patterns, and desire for explainable insights, Logistic Regression is a suitable and defensible choice.

Final Model Metrics: Logistic Regression

Classification Report:

	precision	recall	f1-score	support
0	0.75	0.64	0.69	469
1	0.51	0.64	0.57	281
accuracy			0.64	750
macro avg	0.63	0.64	0.63	750
weighted avg	0.66	0.64	0.64	750

ROC AUC Score: 0.70

Model Summary

The model correctly identifies 64% of customers who are about to churn. While half of the flagged customers may be false alarms, this still allows us to potentially save hundreds of accounts that would otherwise be lost. This trade-off could be acceptable if the cost of churn far outweighs the cost of some extra outreach.

- Class 1 (churn / positive class): Precision is 0.51 and recall is 0.64, indicating the model captures a fair proportion of churners but with a relatively high false-positive rate.
- Balance: The F1 score for churners is 0.57, reflecting this trade-off between precision and recall.

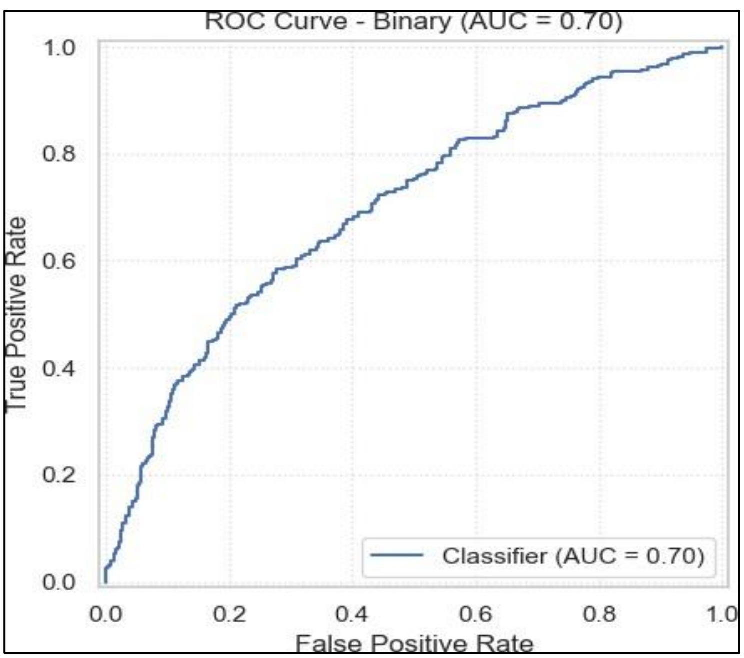
Next Step: Threshold Tuning

Since the model currently balances recall and precision at default thresholds (0.5), tuning the classification threshold will allow us to shift this balance. For churn prediction, recall on the positive class is often more critical, and adjusting the threshold can help capture more at risk customers.

---

5. Model Evaluation

5.1 ROC Curve Summary



The ROC curve shows the model's ability to distinguish between churners and non-churners. With an AUC of 0.70, the model can rank customers in a way that reliably separates likely churners from loyal customers.

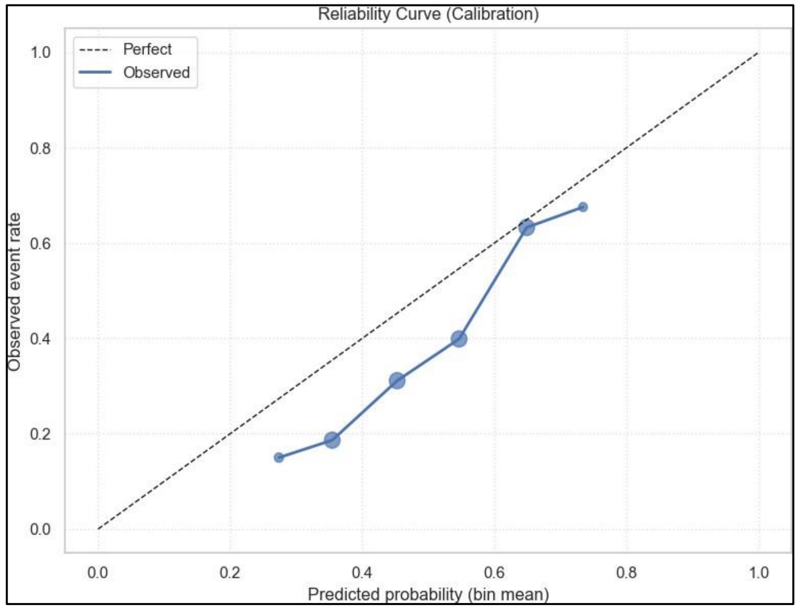
This means retention teams can focus their resources on the riskiest accounts with confidence.

5.2 Calibration

While predictive accuracy is essential, churn modeling requires not just correct classification but also reliable probability estimates. A well-calibrated model ensures that predicted churn probabilities align closely with real-world outcomes.

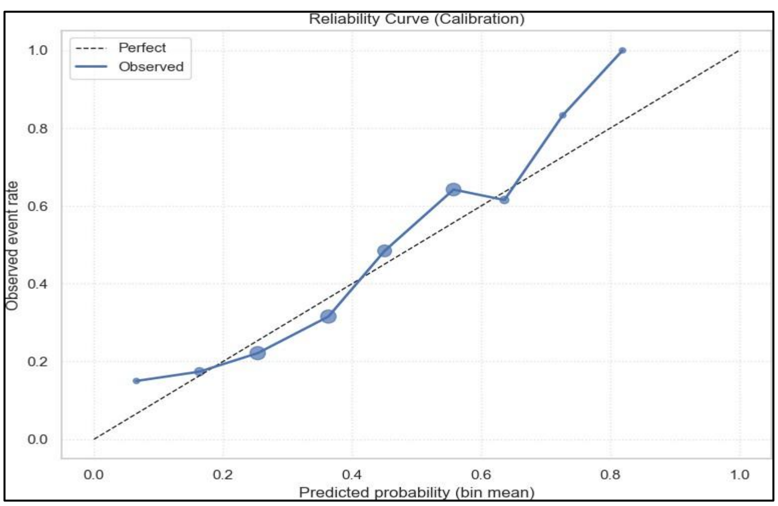
Before Calibration

	Value
Brier Score	0.2243
ECE (Expected Calibration Error)	0.1214



- Brier Score: 0.22 Indicates moderate miscalibration; lower is better.
- Expected Calibration Error (ECE): 0.12 Fairly high, showing that predicted probabilities diverge from actual outcomes.
- Reliability Curve: Points lie below the diagonal, confirming overestimation of churn risk.

The uncalibrated logistic regression tends to predict probabilities that are systematically too high, making it less reliable for risk-based business decisions.



After Calibration

	Value
Brier Score	0.2084
ECE (Expected Calibration Error)	0.0464

- Brier Score: 0.208 Slight Improved (lower error, closer to perfect calibration.
- ECE: 0.0464 Strong improvement, now much closer to ideal well-calibrated.
- Reliability Curve: Much closer to the diagonal, with deviations only at extremes.

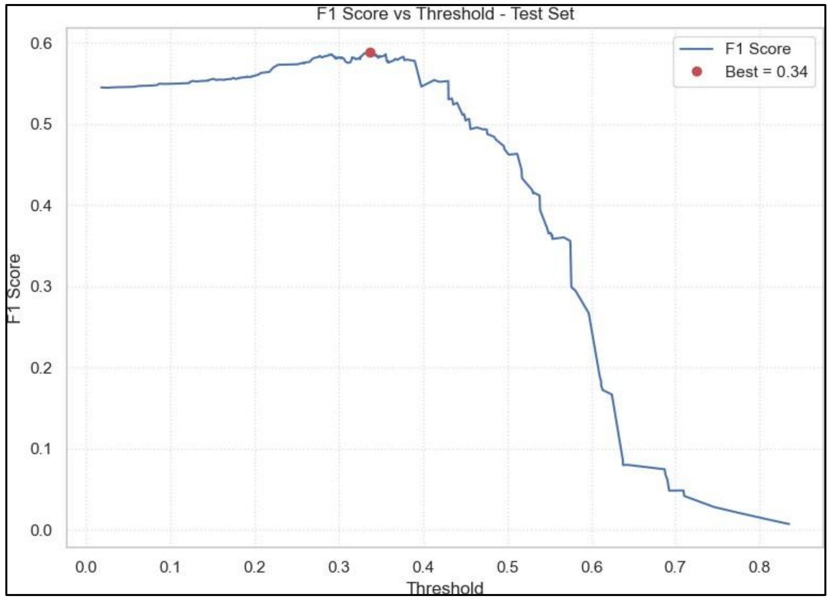
Calibration significantly improved probability estimates, producing predictions that more accurately reflect real-world churn likelihood.

5.3 Model Threshold Tradeoff

F1 Score vs Threshold

This plot shows how the F1 score varies with different classification thresholds. The optimal threshold is identified at 0.34, where the F1 score reaches its peak (~0.59).

This balance point maximizes the trade-off between precision and recall, suggesting that lowering the threshold below the default (0.50) improves the model's ability to capture churners.



Classification Report:

	precision	recall	f1-score	support
0	0.79	0.43	0.56	469
1	0.46	0.81	0.59	281
accuracy			0.58	750
macro avg	0.63	0.62	0.58	750
weighted avg	0.67	0.58	0.57	750

ROC AUC Score: 0.6231779587067205

Summary- Model Performance at 0.45 Threshold

Churn Detection (Class 1):

- Recall: 81% The model correctly identifies a large majority of actual churners, which is essential for proactive retention efforts.
- Precision: 46% About half of the predicted churners are false positives, meaning some customers flagged as at-risk might not churn.

At a threshold of 0.45, the model identifies 8 out of 10 churners, though nearly half of flagged customers may not churn. For the business, this means prioritizing recall ensures fewer customers slip away unnoticed. If the cost of churn outweighs the cost of outreach, this setting maximizes retention value.

6. Business Insights from Model Results

Business Implications

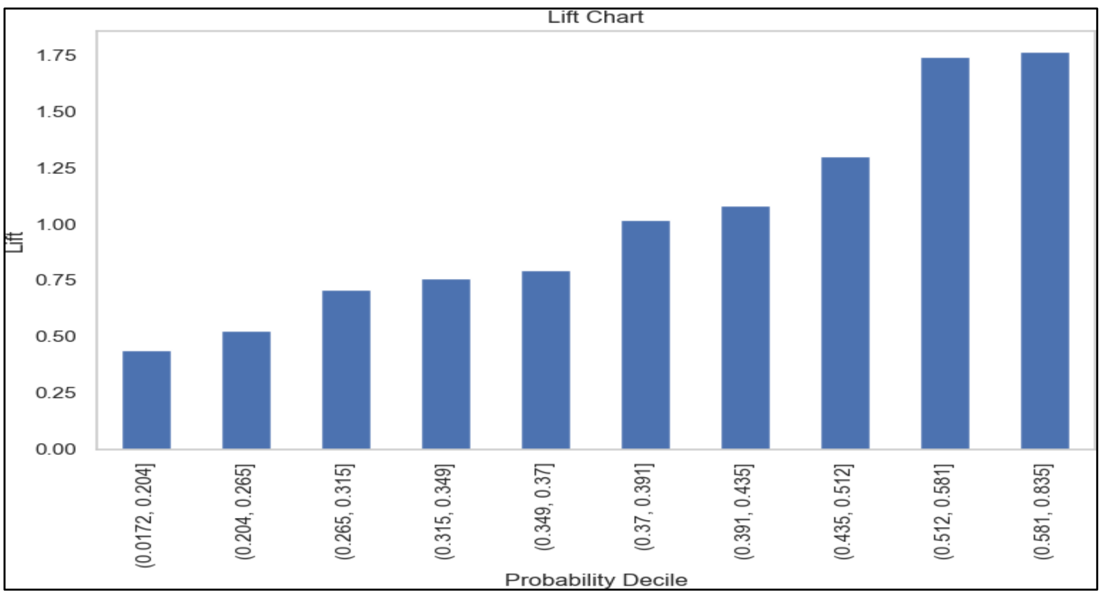
By setting the threshold to 0.45, the model leans conservative, slightly reducing false alarms, which helps:

- Avoid over-targeting customers who are unlikely to churn, saving resources.
- Maintain trust with stable customers who might otherwise receive unnecessary intervention.

However, this comes at the cost of missing some true churners (lower recall for class 0). The choice reflects a business trade-off of targeting fewer but more likely churners.

If the cost of retention offers is high, this threshold setting makes strategic sense. But if every churn matters, a lower threshold favoring recall might be more appropriate.

6.1 Model Lift and Business Value



Lift Chart Summary

The lift analysis shows that the churn model meaningfully separates customers by risk level:

- Top 10% highest-risk customers are 1.8x more likely to churn compared to the overall customer base.
- Churn rates increase steadily across deciles, confirming the model consistently ranks customers in the correct order of risk.
- This segmentation provides a clear roadmap for resource allocation and retention teams can concentrate efforts on the highest-risk deciles, where interventions will deliver the greatest return on investment.

Instead of spreading retention efforts evenly across the customer base, GreenwallTech can now focus on the riskiest 20–30% of accounts, where the probability of churn and the potential revenue at risk is disproportionately high.

6.2 Key Drivers of Churn

	odds_ratio	%_change_in_odds
subscription_type_Basic	1.297867	+29.8%
last_login_days_ago	1.289147	+28.9%
num_support_tickets	1.136197	+13.6%
region_North	1.032686	+3.3%
region_East	1.032603	+3.3%
region_South	0.994628	-0.5%
engagement_score	0.993988	-0.6%
spend_per_tenure	0.989356	-1.1%
age	0.971321	-2.9%
subscription_type_Premium	0.957580	-4.2%
region_West	0.936019	-6.4%
monthly_spend	0.859433	-14.1%
subscription_type_Standard	0.798807	-20.1%
tenure_months	0.788321	-21.2%

Higher Churn Risk Factors

- Basic subscription type customers have about 31% higher odds of churning compared to the baseline (likely Premium or Standard).
- Customers with longer time since last login show ~30% higher odds of churn — indicating disengagement is a strong churn signal.
- A higher number of support tickets is associated with ~14% higher odds of churn, suggesting repeated issues may push customers away.
- Being located in the East or North regions is linked to slightly elevated churn odds (+4–5%), though this effect is modest.

Lower Churn Risk Factors

- Higher monthly spend reduces churn odds by ~14%, pointing to stronger commitment or value perception among higher-spend customers.
- Longer tenure reduces churn odds by ~21%, showing loyalty builds over time.
- Standard subscription type customers have ~20% lower odds of churn than the baseline.
- Customers in the West region have ~7% lower churn odds.

Churn risk is strongly tied to engagement, subscription type, and customer support experience. Retention strategies should prioritize re-engaging Basic plan customers and reducing negative support experiences, especially in the East and North regions. Conversely, retaining high-spend, long-tenure customers should be a lower-cost win since they are already less likely to churn.

8. Business Recommendations & Conclusion

8.1 Business Recommendations

The churn analysis confirms that customer retention at GreenwallTech is driven primarily by engagement, subscription type, and support experience. The predictive model, with an AUC of ~0.70 and recall of 64 - 81% depending on threshold choice, provides a reliable framework for identifying at-risk customers. While not perfect, it offers a defensible and business actionable foundation for targeted retention programs.

Business Recommendations

1. **Prioritize Basic Plan Customers**
   Customers on the Basic plan are 31% more likely to churn. This segment should be the first target of retention campaigns, such as loyalty perks, personalized engagement programs, or upgrades to the Standard plan where churn risk is lower.
2. **Re-Engage Inactive Users**
   Customers with long gaps since last login are significantly more likely to churn. Implementing proactive nudges such as automated reminders, educational content, limited time offers can help re-engage these customers before they fully disengage.
3. **Improve Support Experience**
   High ticket volumes are linked to increased churn. Investing in faster response times and better self-service tools.
4. **Protect High-Value Customers**
   High-spend and long-tenure customers are less likely to churn, but their loss would have a disproportionate financial impact. Proactive account management and early-warning monitoring for this group should help to safeguard recurring revenue.
5. **Regional Focus**
   Customers in the East and North regions show slightly higher churn rates. Localized retention campaigns or region-specific service improvements could help close this gap.

7.2 Model Limitations

- **Potential Confounding Factors**
  The model may not capture all drivers of churn. For instance, external market forces such as competitor pricing and macroeconomic conditions. This limits the model's ability to fully explain churn behavior.
- **Class Imbalance and False Positives**
  Although performance metrics may be acceptable, churn prediction remains imperfect. A relatively high false, positive rate means some retained customers may be incorrectly flagged as at risk, potentially leading to inefficient retention efforts or customer dissatisfaction.
- **Model Assumptions**
  Logistic regression assumes linear relationships between predictors and churn probability. While exploratory analysis suggested this was reasonable, nonlinear effects could exist. ***More complex models with gridsearch for hyperameter tuning, may be considered in the future*** to capture subtle patterns missed by the chosen model.
- **Calibration & Business Use**
  While calibration improved probability estimates, churn probabilities remain approximations. Business users should treat these scores as risk indicators, not certainties. Intervention strategies should be tested in controlled pilots before scaling.

Final Note

The churn model will require ongoing monitoring, recalibration, and retraining as customer behavior evolves. To maximize impact, pilot retention campaigns should be tested with small customer groups, then scaled once proven effective. By combining predictive insights with targeted interventions, GreenwallTech can meaningfully improve retention rates, protect revenue, and strengthen long-term customer relationships.