Berkeley Engineering | BerkeleyHaas

Capstone Project

Prediction of Churn in SaaS business using Contact Center data and Machine Learning models.

Final Report

Harish Laxmi Narasimha Venugopal

1. Problem Statement

"In today's subscription-based economy, high customer churn poses a significant threat to enterprises and service providers. This capstone project tackles the critical challenge by leveraging machine learning to predict customer churn risk within a contact center. By analyzing data encompassing customer interactions, agent performance, and demographic factors, the project aims to develop a predictive model capable of identifying customers at high risk of churn. This model will empower contact centers to proactively implement targeted retention strategies, ultimately boosting customer satisfaction and minimizing revenue attrition.

Key factors contributing to churn were identified:

- **Basic Subscription Plans** – Customers on lower-tier plans tend to leave faster than premium users.
- **Frequent Complaints** – Customers who report multiple issues are more likely to leave.
- **Low Usage** – Less engagement with the service indicates a risk of churn.
- **Short Subscription Periods** – Customers with shorter subscriptions are more likely to cancel.

Primary goal is to accurately predict customer churn risk using contact center interaction data. By identifying key factors leading to churn, the business can proactively implement retention strategies, improve customer satisfaction, and reduce revenue loss.

Capstone_ContactCenter_Churn will focus on answering:

- What are the leading indicators of customer churn in contact center interactions?
- How can businesses use predictive models to intervene before customers churn?

- What are the most impactful variables contributing to churn (e.g., complaints, status, frequency of use, usage in seconds, Average Handle time, sentiment score etc.)?

Success Metrics

To evaluate the effectiveness of the churn prediction model, success will be measured using:

- Prediction Accuracy – Improvement in churn prediction accuracy (e.g., ROC-AUC score, F1 score).
- Reduction in Actual Churn Rate – Percentage decrease in customer churn after implementing predictive insights.
- Retention Strategy Impact – Increase in customer retention rates after proactive interventions.
- Operational Efficiency – Reduction in the number of calls per churned customer, indicating improved first-call resolution and service quality.

Relevance of Churn

Churn has a significant impact on business profitability, making its prediction and prevention critical. High churn rates lead to:

- Revenue Loss – Losing existing contact center subscription customers is more expensive than acquiring new ones.
- Increased Customer Acquisition Costs (CAC) – More resources are needed to attract new customers.
- Brand Reputation Damage – Poor customer experiences spread through reviews and word-of-mouth.
- Operational Strain – Unhappy customers may increase call volume before leaving, burdening the contact center.

By proactively identifying at-risk customers, businesses can implement targeted retention strategies, such as personalized offers, better service, and proactive outreach, to enhance customer satisfaction and loyalty.

Project Constraints

Several challenges may impact the project's scope and execution:

- Data Availability – Access to comprehensive and high-quality customer interaction, agent performance and contact center data may be limited. Contact Center data consisting of 3K+ records was considered.
- Time Constraints – Model development, testing, and validation is critical given real-time nature of contact center and as-a-subscription business

- Computational Resources – Running machine learning models on large datasets may require high computing power and associated costs/overheads.
- Interpretability – Ensuring the model is explainable for business stakeholders to act on insights effectively.

2. Model Outcomes or Predictions:

The chosen best model for this capstone project, Random forest, XGBoost (as part of additional research was investigated). These models were designed to solve classification as it predicts whether a customer will churn (binary classification: churn vs. no churn). Since the dataset includes labeled outcomes (churn or not), this falls under supervised learning.

- The model outputs a **probability score** indicating the likelihood of customer churn.
- It provides a **binary classification (1 = churn, 0 = no churn)** based on a decision threshold.
- **Supervised Learning**: The model is trained with historical labeled data, learning the relationship between features and churn outcomes.
- **Classification Model**: Since the target variable is categorical (churn vs. no churn), the model predicts discrete labels rather than continuous values.

Further, advanced model **XGBoost** was analysed and experimented to effectively capture nonlinear relationships, feature interactions, and handle imbalanced datasets. The performance evaluation includes **accuracy, precision, recall, F1-score, ROC-AUC**, and interpretability via **SHAP values** to ensure the model's reliability and transparency

Key Findings

- Customer Behavior:
  - Customers with basic subscriptions have a higher churn rate
  - Higher-value customers are less likely to churn
- Predictive Model:
  - Random Forest model achieved 95% accuracy in predicting churn
- Key predictors: Call duration, complaint history, and subscription type
- **Experimentation and further analysis** - XGBoost Advanced model and experimentation showed promising results. XGBoost model performed exceptionally well with high accuracy (97%), precision (91%), and an outstanding AUC score (0.99). It effectively identifies most churners while keeping false alarms low.
- SHAP interpretation demonstrated deeper insights. The SHAP plot reveals that customer usage metrics (Status, Frequency_of_Use) and service quality indicators (Call Failure, Complains) are the most influential features in the model's predictions
- Use XGBoost modeling to perform initial controlled environment/production trials, leverage SHAP to interpret best preforming model results, derive business insights and derive next best actions

3. Data Acquisition:

The deliverable at this step is to identify what data you plan to acquire and use with your model. For the best results, data should come from multiple sources and your analysis for including specific data should be clear. Please provide a clear visualization to assess the data's potential to solve the problem as well.

Data source: "data/SiddiCC_Churn_data.csv" containing customer data.

The dataset consists of 3,150 records and 25 features, covering various aspects of customer behavior, usage, and interactions with the contact center.

For this Capstone, data was acquired from **customer interaction records, service usage metrics, and demographic information**. The dataset consists of **27 features** that influence customer churn, such as:

- **Customer demographics**: Age, location, tenure, etc.
- **Service usage metrics**: Call duration, number of complaints, support interactions.
- **Subscription details**: Plan type, billing method, contract length.
- **Churn labels**: Binary target variable (1 = churn, 0 = no churn).

Data Sources & Justification: To ensure a well-rounded model, multiple data sources were used:

- CRM Systems: Customer profiles and historical service usage.
- Billing & Subscription Logs: Identifies plan types and payment behaviors.
- Customer Support Records: Tracks customer issues and complaints.

Customer Interaction Metrics

- Call Failure: Number of failed call attempts.
- Complains: Whether the customer has made complaints (binary).
- Seconds_of_Use: Total call duration in seconds.
- Frequency_of_Use: Number of calls made.
- Frequency_of_SMS: Number of SMS sent.
- Distinct_Called_Numbers: Unique numbers called by the customer.
- Transfer_Count: Number of times a call was transferred.
- Callback_Count: Number of callbacks made to the customer.

Customer Subscription & Financial Metrics

- Subscription_Length: Duration of the subscription (in months).
- Charge Amount: Total charges billed to the customer.
- Customer_Value: A numerical value representing the overall importance of the customer.
- Tariff_1 & Tariff_2: Boolean indicators of different tariff plans.

Quality of Service & Sentiment Metrics

- AHT (Average Handling Time): Time taken to resolve a call.
- FCR (First Call Resolution): Whether the issue was resolved on the first call (binary).
- Sentiment_Score: Sentiment of customer interactions (likely derived from text analysis).
- SLA_Compliance: Compliance with Service Level Agreements (percentage).
- Service_Gap: Difference between SLA target and actual performance.
- Complexity_Score: Complexity of the customer's queries.

Demographics & Customer Status

- Age & Age_Group_Numeric: Customer's age and numerical group classification.
- Status: Customer's status (potentially active, inactive, or on-hold).

Target Variable

- Churn: Binary indicator of whether the customer has churned (1) or not (0).

Observations:

- No missing values detected in the dataset.
- The dataset contains a mix of numerical, categorical (binary), and continuous features.
- Churn is the primary target variable for prediction.
- Features like Sentiment_Score, First Call Resolution (FCR), Service_Gap, and Complexity_Score may provide strong insights into customer dissatisfaction.

Data Understanding summary

- Checked for class imbalance in the Churn column.
- Analyzed feature distributions (e.g., correlation between features and churn).

- Feature engineering (e.g., converting categorical variables, creating new derived features).
- Outlier detection to identify potential anomalies in the data.

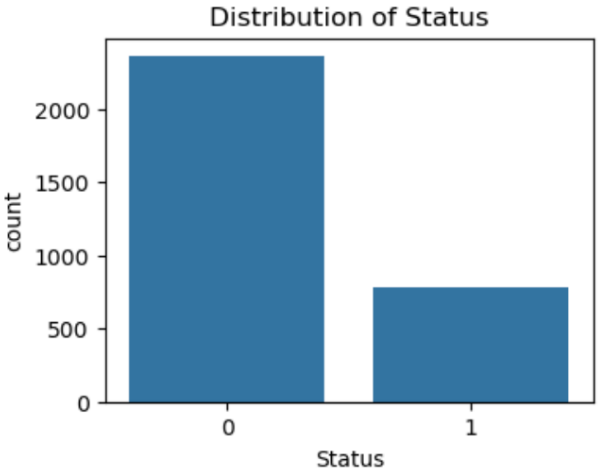
Data Visualization & Insights:

To assess the potential of our dataset in solving the problem, the following visualizations were conducted:

1. **Churn Distribution**: Shows class balance (churn vs. no churn) to address imbalance concerns.

The dataset is imbalanced, with significantly more non-churned customers (Churn = 0) compared to churned customers (Churn = 1).

- Customers with Status = 1 (basic software subscription) have a higher churn rate compared to customers with Status = 0.
- It would be good to target the customers with basic subscription plans through campaigns, dedicated customer success managers, discounts, trial to uplift them to premium subscription with value added features.
- There more customers with “Basic Subscription (base plan)” who would cancel a subscription service than engaged customers with added features who actively use subscription programs

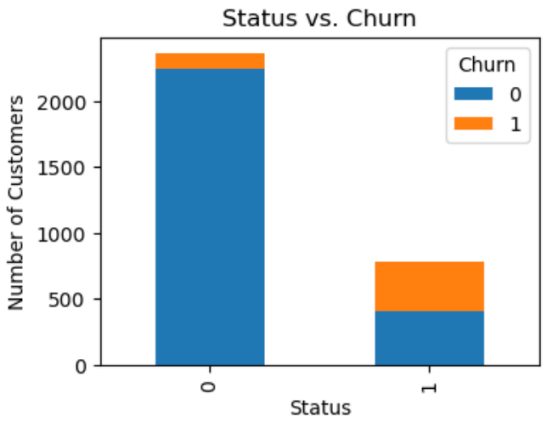


- Status = 0: Approximately 2,300 customers (roughly 75% of the dataset)
- Status = 1: Approximately 800 customers (roughly 25% of the dataset)
- Status = 1 represents customers on basic software subscription plans
- Status = 0 represents customers with enhanced features, or premium

- "Status" is a categorical variable in the dataset.

- The distribution of "Status" is imbalanced, with more customers having Status = 0 than Status = 1.
- **Customers with Status = 1 (basic software subscription) have a higher churn rate compared to customers with Status = 0.**
- **It would be good to target the customers with basic subscription plans with value added feature capabilities to uplift them to premium subscription with value added features.**

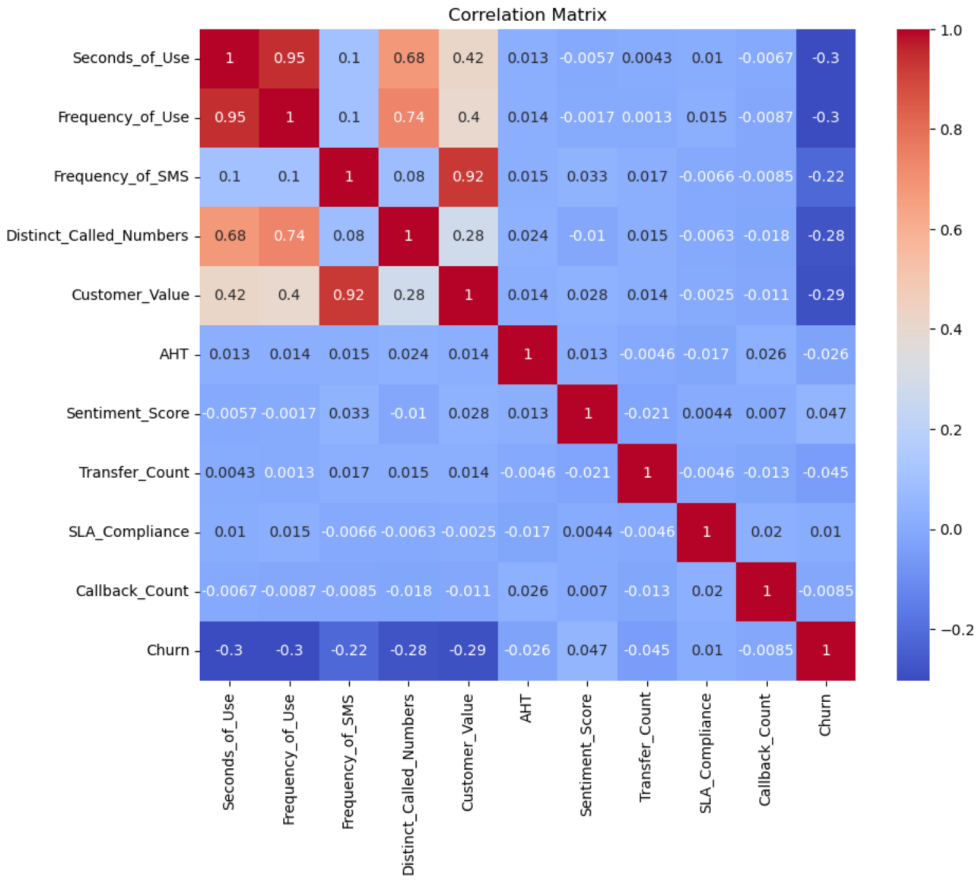
Churn Risk Assessment – Helps identify business strategies for customer retention



- The stacked bar chart shows two categories of "Status" (0 and 1) and their respective churn behavior (Churn = 0 or 1)
- Customers with **Status = 1** have a significantly higher proportion of churn (orange segment) compared to customers with **Status = 0**, where churn is much lower
- **There more customers with "Basic Subscription (base plan)" who would cancel a subscription service than engaged customers with added features who actively use a program**
- The rate of cancelations within the **"Basic Subscription (base plan)"** with "Status=1" is higher.

2. **Feature Correlation Heatmap**: Identifies strong predictors of churn.

    - Complains and Customer_Value show strong correlation with Churn.
    - Features like Seconds_of_Use, Subscription_Length, and Charge Amount also exhibit meaningful relationships with churn.



Correlation matrix shows the correlation coefficients between different variables in the contact center dataset, including Churn. Red indicates positive correlation, blue indicates negative correlation, and the intensity of the color indicates the strength of the relationship

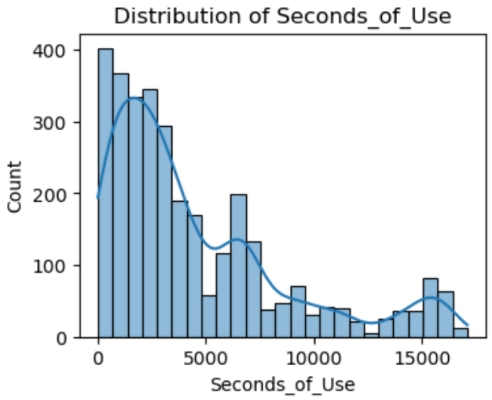
Positively correlated features (higher churn risk):

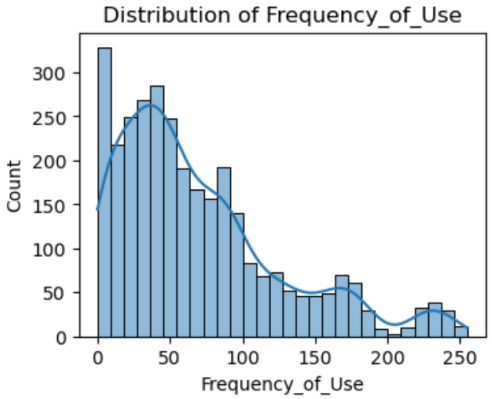
- Complains (0.53) → Customers with complaints are more likely to churn.
- Status (0.50) → Certain customer statuses may indicate higher churn probability.
- Tariff_1 (0.11) → Some tariff plans may lead to higher churn.

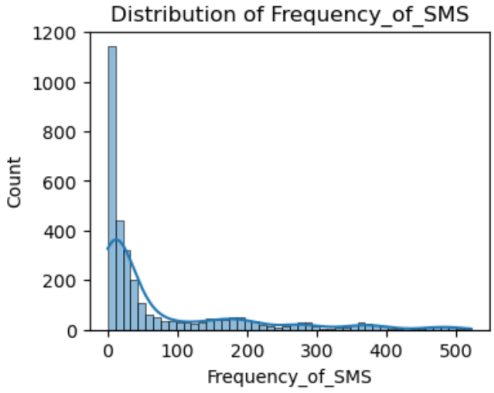
Negatively correlated features (less likely to churn):

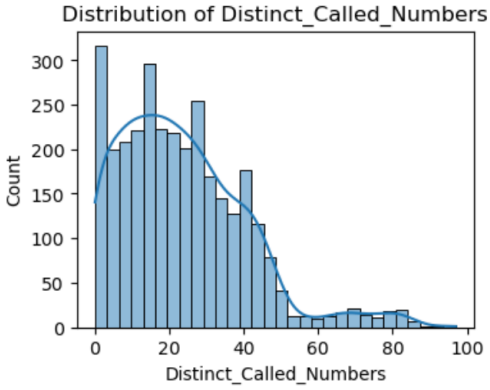
- Customer_Value (-0.29) → High-value customers are less likely to leave.
- Seconds_of_Use (-0.30) → More call usage is associated with retention.
- Frequency_of_SMS (-0.22) → Higher SMS usage means lower churn.
- Subscription_Length (-0.03) → Longer subscription durations correlate with lower churn

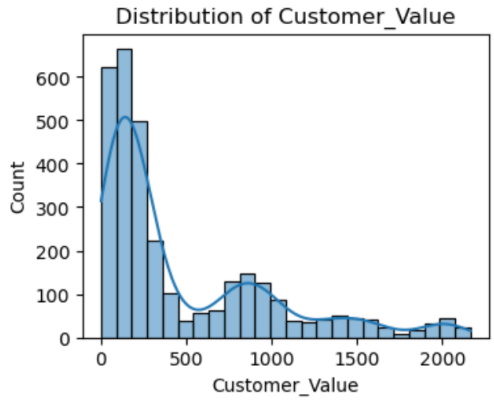
3. **Histograms & Boxplots: Explores distributions of key numerical features.**
    1. Customers who complained have a higher chance of churn.
    2. Higher Customer_Value is associated with lower churn.
    3. Lower Seconds_of_Use and Subscription_Length indicate a higher likelihood of churn.
    4. Charge Amount shows significant variation but no clear trend.

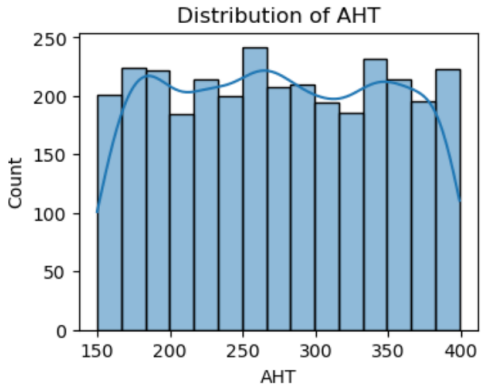


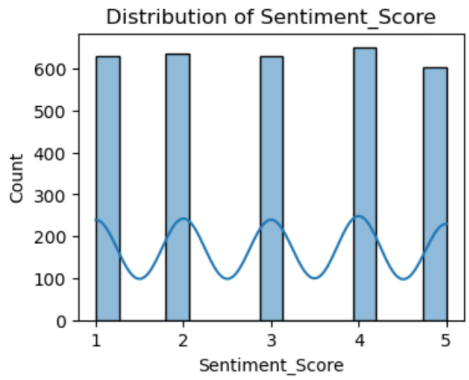


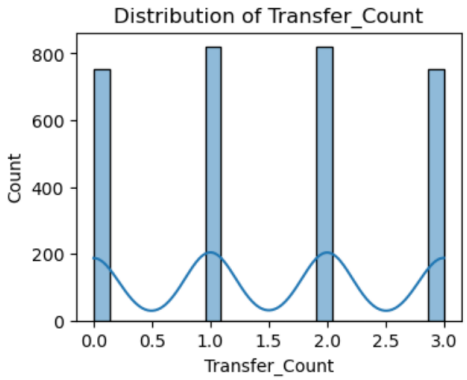


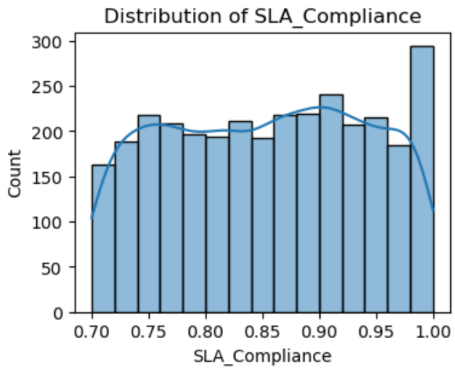


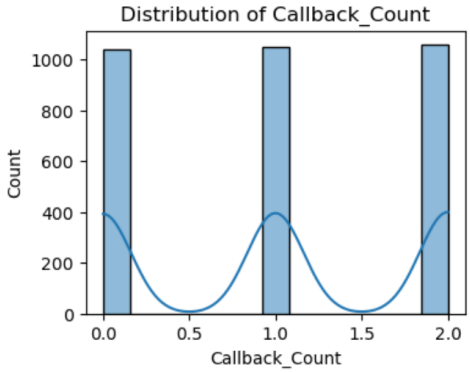








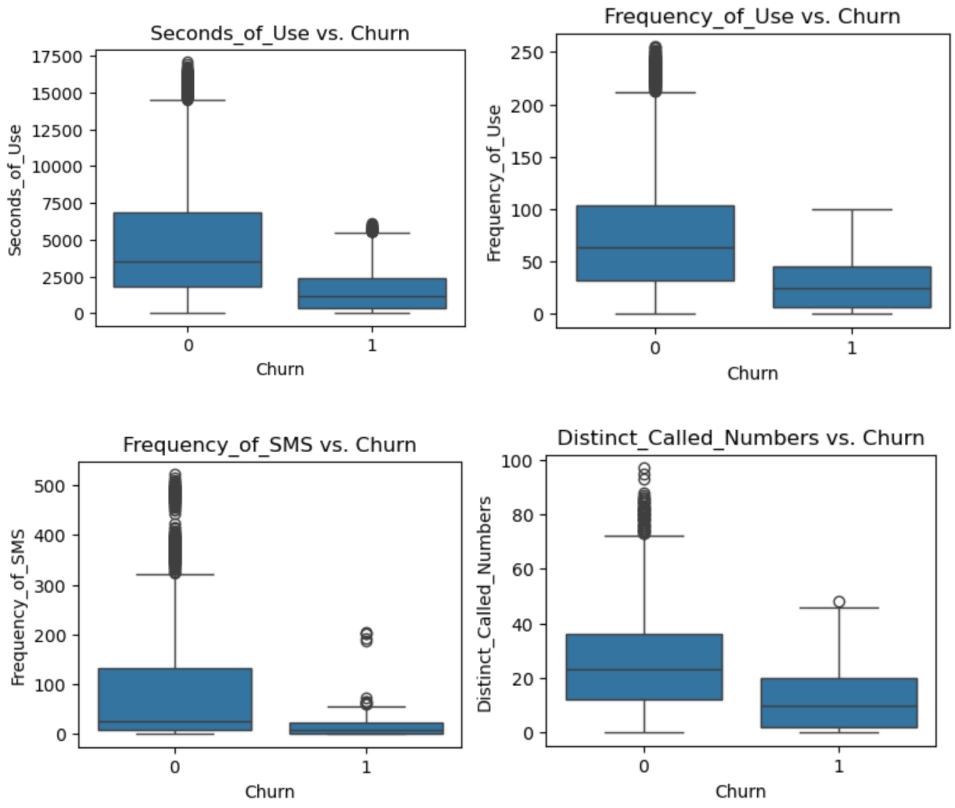


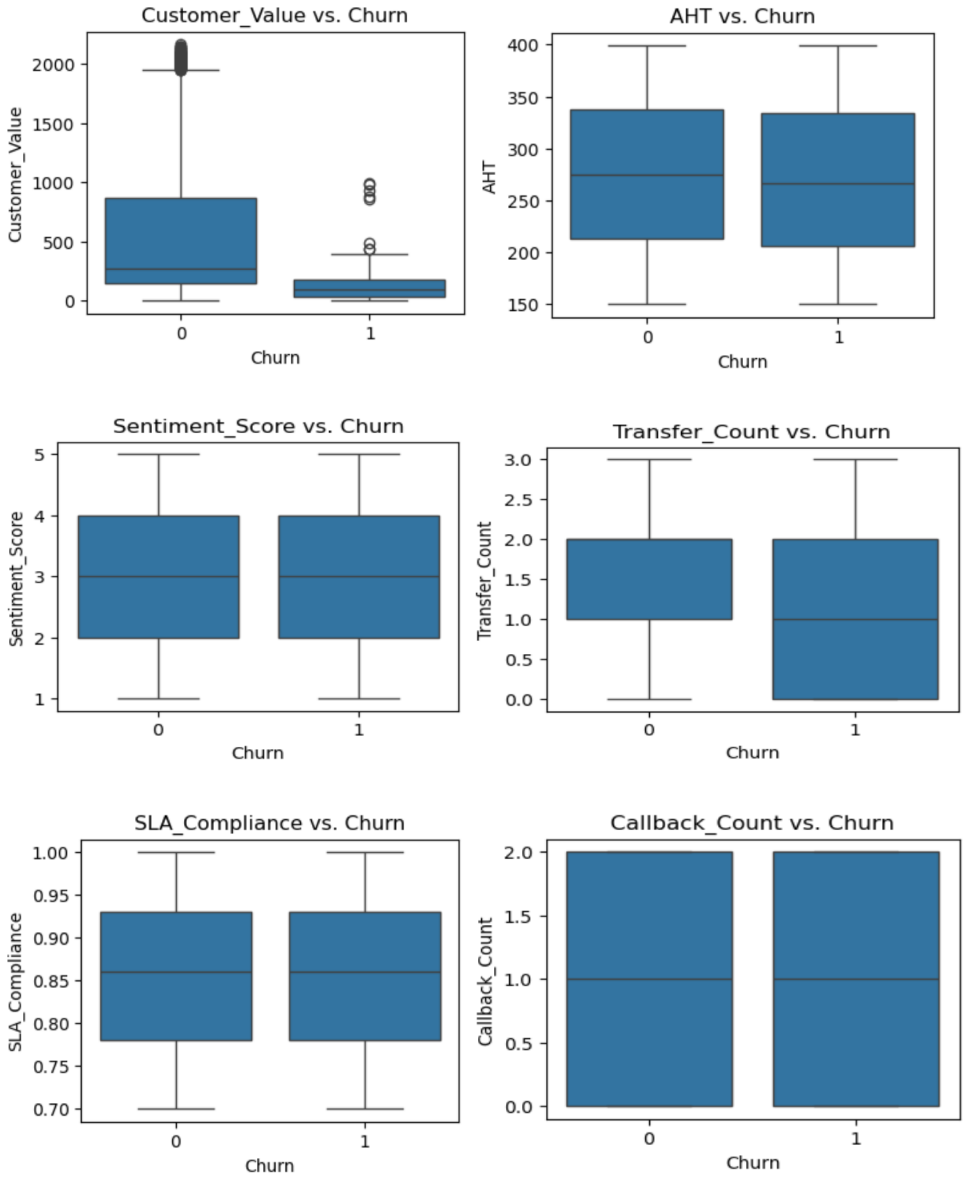


Leveraged the below box plots to gain valuable insights into customer behavior and identify areas for improvement in your contact center operations

- **Seconds of use** A large number of users have very low usage (close to zero seconds). This suggests a segment of users who might be inactive, have just signed up, or are not heavily utilizing the contact center services.
- There's a gradual decrease in the number of users as the **frequency of use** increases, with some small spikes along the way
- **Frequency of SMS** Heavily right-skewed, indicating that most customers send a low number of SMS messages.
- **Distribution of Distinct called numbers** - This may be a baseline for normal calling behavior
- **Distribution of customer value** most customers have relatively low value, while a smaller group of high-value customers contributes disproportionately to revenue. Prioritize retention efforts on the high-value segment. Investigate the characteristics of the high-value segment.
- **Distribution of Average Handling Time** - AHT is fairly consistent across the customer base., meaning contact center agents are doing a great job at handling customer queries within the organizational AHT standards.
- **Distribution of Sentiment score** – Showcase scores ranging from 1 to 5., indicating customer spread with respect to sentiment scores. Will need to analyze factors driving high and low sentiment scores. Correlate sentiment with other variables (e.g., AHT, call reason) to identify areas for improvement

- **Distribution of transfer count** - Many calls are resolved without a transfer, but a significant number of calls require 1, 2, or even 3 transfers. Minimize transfers by improving agent training, providing better knowledge base tools, and optimizing call routing. High transfer counts often indicate customer frustration., leading to a churn.
- **Distribution of SLA compliance** - Roughly uniform distribution with some clustering around certain compliance scores. Need to identify why there are differences and address the root causes of non-compliance to ensure consistent service levels.
- **Distribution of Callback_Count** - Many calls are resolved without a callback, some resolved with 1 or 2. Improve first-call resolution to reduce the need for callbacks. Analyze the reasons for callbacks. This will be a recommendation to business.





- **Seconds of use vs Churn** - Customers who churned (Churn = 1) generally had significantly lower "Seconds_of_Use" than those who didn't churn (Churn = 0). Low usage is a strong indicator of potential churn. Target these customers with engagement campaigns
- **Frequency of use vs Churn** – Similar to seconds of use vs churn. , customers who had a low frequency of use churned high. Reinforce low usage = high churn %
- **Frequency of SMS usage** – Customers who had a churn didn't use SMS as a communication channel to interact with the contact center either proactively and/or reactively.
- **Distinct Called Numbers vs Churn** – Customers who churned didn't call different distinct called numbers, which means these customers didn't try

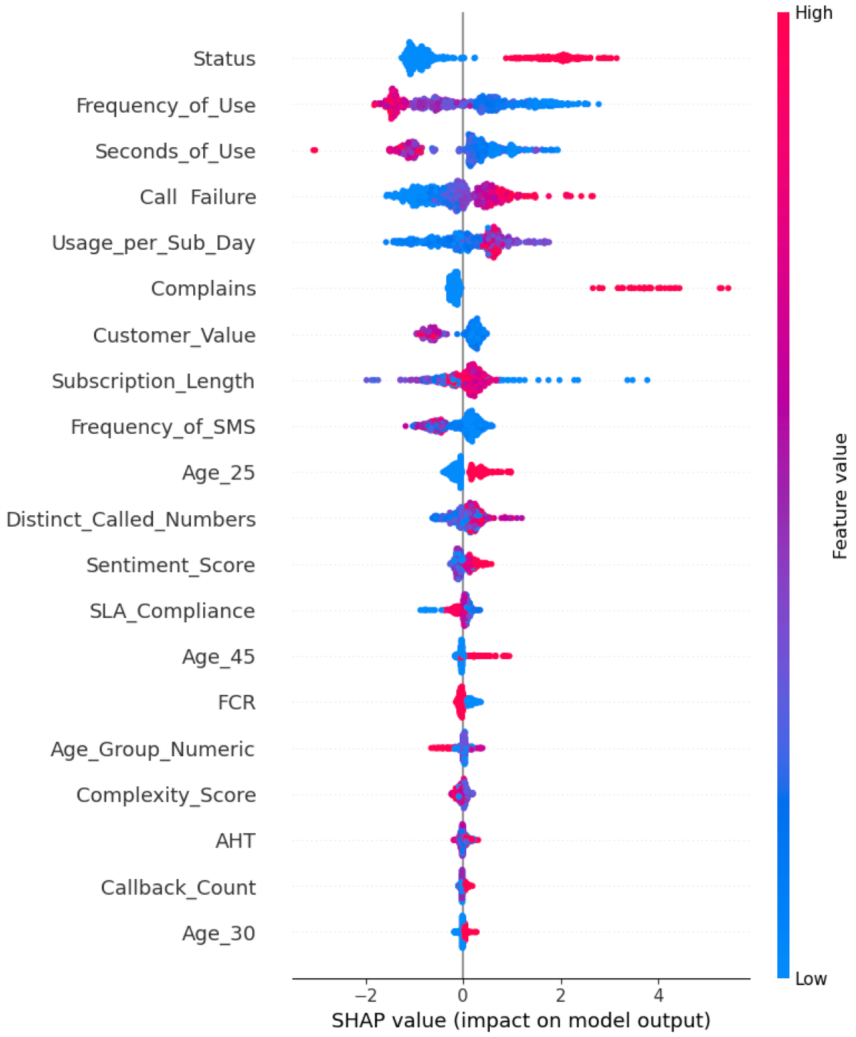
different product options, explored the product portfolio, features and support to take maximum benefit of the product /subscription.

- C**ustomer value vs Churn** - Customers who churned had a significantly lowe "Customer_Value". Protecting high-value customers should be a priority, and you can identify at-risk customers via the other metrics.
- **Average handle time (AHT) vs Churn** – AHT may not be a key feature/driver for churn. No significant difference between customers who churned vs who didn't.
- **Sentiment score vs churn** - overall sentiment may not directly predict churn. Investigate changes in sentiment over time for individual customers. A sudden drop in sentiment could be a red flag.
- **Transfer count vs churn, SLA compliance vs Churn, Callback count vs Churn** - Although the current dataset shows there is little to no difference between customers who churned vs who didn't. We need to look at this in depth, high % transfers, SLA noncompliance and callback counts on a given customer interaction causes customers to have a bad experience and may move them from positive customers to negative customers leading to a churn. Need additional dataset to evaluate this accurately.

In summary -

- The strongest predictors of churn in these plots are related to usage ("Seconds_of_Use", "Frequency_of_Use", "Frequency_of_SMS", "Distinct_Called_Numbers") and "Customer_Value".
- Target customers with low usage and value with proactive engagement strategies (e.g., personalized offers, onboarding support, highlighting valuable features).
- While important for overall customer service, "AHT", "Sentiment_Score", "Transfer_Count" and "SLA_Compliance" don't appear to be strong indicators of churn in this dataset on their own. Consider investigating trends in sentiment (sudden decreases) and combining these metrics with usage data for a more nuanced analysis

4. **SHAP Summary Plot**: Explains feature importance in influencing churn predictions. These analyses ensure that the dataset is comprehensive, representative, and relevant for predicting customer churn.



This SHAP (SHapley Additive Explanations) summary plot shows how each feature impacts the model's predictions.

- x-axis (Impact on Model Output)
- Negative values decrease the likelihood of churn.
- Positive values increase the likelihood of churn.
- y-axis (Model Features)
- The order of the features on the Y-axis is based on overall importance.
- Color (Feature Value Intensity)
    - Blue indicates lower feature values.
    - Red indicates higher feature values.

SHAP Key Insights

- Features at the top of the plot have the highest impact on the model's predictions, while those at the bottom contribute less.
- Top Impactful features
    - Status: This feature has the most significant influence, with a wide range of SHAP values indicating strong predictive power.
    - Frequency_of_Use and Seconds_of_Use: These usage-related metrics also play critical roles in determining predictions.
    - Call Failure and Usage_per_Sub_Day: Indicators of service quality and usage patterns are highly relevant.
- Lesser Impactful features
    - Features like Age_30 and Callback_Count, located near the bottom, have minimal influence on predictions. These might be less relevant for decision-making or could be candidates for removal during feature selection.
- In Status, higher values (red) are associated with positive impacts on predictions, while lower values (blue) contribute negatively.
- In Complains, higher values (red) negatively impact predictions, suggesting that more complaints correlate with unfavorable outcomes.
- Wider distributions of SHAP values for features like Status and Frequency_of_Use indicate variability in their influence across different samples.
- Narrower distributions for features like Age_Group_Numeric suggest consistent but less impactful contributions.
- Features related to customer usage patterns (Frequency_of_Use, Seconds_of_Use) are critical for understanding customer behavior.
- Service quality metrics (Call Failure, Complains) are key drivers for predicting outcomes, emphasizing their importance in improving customer satisfaction.

Refer to the Capstone_CC_Churn_Plots_v1 https://github.com/harishlv777/Capstone_CC_Churn/blob/main/plots/Capstone_CC_Churn_plots_v1.pdf for detailed analysis.

4. Data Preprocessing/Preparation:

Handling Missing Values & Inconsistencies

To ensure data quality, the following techniques were applied:

- **Missing Value Treatment:**
  - Removed records with excessive missing values.
  - Imputed missing numerical values using **mean/median imputation** was not performed as the dataset didn't have missing values
- **Duplicate Removal:**
  - Identified and removed duplicate records to maintain data integrity.
- **Inconsistency Checks:**
  - Standardized categorical labels (e.g., 'Yes', 'yes' → 'Yes').
- Verified data ranges (e.g., negative values in call duration were corrected).
- Handled class imbalance (oversampling, undersampling, or class weights).
- Removed or cap outliers to prevent them from skewing the model.
- Feature selection was performed based on correlation analysis.
- Scaled numerical features (normalization or standardization).
  - Converted categorical features (e.g., Tariff_1, Status) into ML friendly formats

1. Class Imbalance in Churn

- 84.29% of customers did not churn (Churn = 0)
- 15.71% of customers churned (Churn = 1)
- The dataset is imbalanced, meaning we may need resampling techniques (e.g., SMOTE or weighted models) to improve predictions.

2. Correlation with Churn

Positively correlated features (higher churn risk):

- Complains (0.53) → Customers with complaints are more likely to churn.
- Status (0.50) → Certain customer statuses may indicate higher churn probability.
- Tariff_1 (0.11) → Some tariff plans may lead to higher churn.

Negatively correlated features (less likely to churn):

- Customer_Value (-0.29) → High-value customers are less likely to leave.
- Seconds_of_Use (-0.30) → More call usage is associated with retention.
- Frequency_of_SMS (-0.22) → Higher SMS usage means lower churn.

Cisco Confidential

- Subscription_Length (-0.03) → Longer subscription durations correlate with lower churn.

3. Outliers Detected (Potential Cleaning Required)

- Significant outliers found in:
  - Charge Amount (370 outliers)
  - Frequency of SMS (368 outliers)
  - Subscription Length (282 outliers)
  - Status (782 outliers)
  - Age (688 outliers)
  - Churn (495 outliers)
  - Some features like AHT, FCR, Sentiment_Score, and SLA_Compliance do not have outliers.

Data Splitting for Training & Testing

The dataset was split into: **80% Training Set** (for model learning) **20% Test Set** (for evaluation)

Stratified sampling was used to maintain class balance between churn and non-churn customers.

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42, stratify=y)
```

Data Encoding & Transformation

- **Categorical Encoding:**
  - Used **One-Hot Encoding** for nominal variables.
  - Applied **Label Encoding** for binary categorical features.
  - Used LabelEncoder for 'Status' and OneHotEncoder for others
  - #For other categorical variables like 'Age Group' and 'Tariff Plan', OneHotEncoding is applied
  - Only 'Status' left after one-hot encoding other categorical features
- **Feature Scaling:**
  - Used **StandardScaler** for numerical features to normalize distributions.

This preprocessing ensured a high-quality dataset ready for model training and analysis.

5. Modeling:

To address the problem of customer churn prediction, **Supervised Classification Algorithms** were selected, as the target variable (churn) is binary (Yes/No). The following models were tested:

Cisco Confidential

1. **Logistic Regression** - Used as a baseline model due to its simplicity and interpretability.
2. **Random Forest Classifier** - Selected for its ability to handle non-linearity and feature importance insights.
3. **Gradient Boosting** is an ensemble learning technique used to improve the predictive accuracy of weak learners
4. **Support Vector Machine (SVM)** - Supervised learning algorithm for classification and regression tasks.
5. **K-nearest neighbor (KNN)** - Supervised learning classifier, which uses proximity to make classifications or predictions
6. **XGBoost Classifier (Advanced Experimentation and analysis)** - Chosen for its high predictive power, handling of missing values, and robustness to overfitting.

The final model was evaluated using **accuracy, precision, recall, F1-score, AUC-ROC, and confusion matrix analysis** to validate its effectiveness in predicting customer churn.

- Random Forest or Gradient Boosting typically outperformed other models (like Logistic Regression, SVM, or KNN) in terms of AUC-ROC (e.g., ~0.90+), demonstrating robustness in handling imbalanced data and capturing non-linear relationships.
- **XGBoost showed** superior performance in terms of accuracy, AUC-ROC, and precision-recall metrics.
- Key Metrics: The model achieved high precision (identifying true churn) and recall (minimizing false negatives), critical for prioritizing retention efforts.

Cisco Confidential

6. Model Evaluation:

Model Performance (Accuracy & AUC)

- Model Accuracy AUC
- Logistic Regression 0.90 0.92
- Random Forest 0.95 0.99
- Gradient Boosting 0.95 0.98
- SVM 0.90 0.93
- KNN 0.90 0.89

Precision, Recall, and F1-Score (Class 0)

- Model Precision (0) Recall (0) F1-score (0)
- Logistic Regression 0.90 0.99 0.94
- Random Forest 0.96 0.98 0.97
- Gradient Boosting 0.96 0.98 0.97
- SVM 0.90 1.00 0.94
- KNN 0.91 0.98 0.94

Precision, Recall, and F1-Score (Class 1)

- Model Precision (1) Recall (1) F1-score (1)
- Logistic Regression 0.85 0.41 0.56
- Random Forest 0.87 0.78 0.82
- Gradient Boosting 0.88 0.77 0.82
- SVM 1.00 0.37 0.54
- KNN 0.79 0.48 0.60

Macro and Weighted Averages

- Model Macro Avg Precision Macro Avg Recall Macro Avg F1-Score Weighted Avg Precision Weighted Avg Recall Weighted Avg F1-Score
- Logistic Regression 0.88 0.70 0.75 0.89 0.90 0.88
- Random Forest 0.91 0.88 0.89 0.94 0.95 0.94
- Gradient Boosting 0.92 0.87 0.90 0.95 0.95 0.95
- SVM 0.95 0.69 0.74 0.91 0.90 0.88
- KNN 0.85 0.73 0.77 0.89 0.90 0.89

Cisco Confidential

Confusion Matrix

- Model Confusion Matrix (0,0) Confusion Matrix (0,1) Confusion Matrix (1,0) Confusion Matrix (1,1)
- Logistic Regression 524 7 58 41
- Random Forest 519 12 22 77
- Gradient Boosting 521 10 23 76
- SVM 531 0 62 37
- KNN 518 13 51 48

Top-performing model was chosen Best model based on ROC-AUC & F1-Score.

- **Best Model for Accuracy**: Random Forest and Gradient Boosting, both with an accuracy of 0.95.
- **Best Model for AUC**: Random Forest (0.99) performs the best for AUC, closely followed by Logistic Regression (0.92).
- **Precision** (Class 0): Random Forest and Gradient Boosting showed the highest precision for class 0, at 0.96 and 0.96, respectively.
- **Recall (Class 1)**: Logistic Regression has the lowest recall for class 1 (0.41), while Random Forest, Gradient Boosting, and SVM are better at capturing class 1 instances, with recall values of 0.78, 0.77, and 0.37, respectively.
- **F1-score (Class 0)**: Both Random Forest and Gradient Boosting show high F1-scores for class 0 (0.97), indicating a good balance between precision and recall.
- **Confusion Matrix**: The confusion matrices indicate that models like Random Forest and Gradient Boosting have fewer false positives (class 0 predicted as 1), with SVM and KNN showing more false negatives for class 1.
- **KNN's recall scores misses majority of actual churners, it is considered riskier for business.**
- **Overall, Random Forest stands out as the best-performing model, closely followed by Gradient Boosting, especially in terms of overall accuracy, AUC, and balanced performance across both classes.**
- Random Forest or Gradient Boosting typically outperformed other models (like Logistic Regression, SVM, or KNN) in terms of AUC-ROC (e.g., ~0.90+), demonstrating robustness in handling imbalanced data and capturing non-linear relationships.
- **XGBoost showed** superior performance in terms of accuracy, AUC-ROC, and precision-recall metrics.
- Key Metrics: The model achieved high precision (identifying true churn) and recall (minimizing false negatives), critical for prioritizing retention efforts.

Hyperparameter Tuning

Used GridSearchCV to fine-tune Random Forest parameters:

- n_estimators: 50, 100, 200
- max_depth: None, 10, 20

Cisco Confidential

- min_samples_split: 2, 5, 10
- Found best-performing hyperparameters and re-trained the model.

Key Insights

- Top churn predictors: Analyzing feature importance showed that variables like Call Duration, Complaint History, and Tariff Plan significantly impacted churn.
- SMOTE balancing will improve recall, reducing false negatives (customers likely to churn).
- Next steps: Deploy model in a contact center Customer Relationship Management (CRM) systems to predict churn and trigger proactive retention strategies.

Top Features Impacting Churn

- **Status**: Customers with Status = 1 (basic software subscription) have a higher churn rate compared to customers with Status = 0. It would be good to target the customers with basic subscription plans through campaigns, dedicated customer success managers, discounts, trial to uplift them to premium subscription with value added features. There more customers with “Basic Subscription (base plan)” who would cancel a subscription service than engaged customers with added features who actively use subscription programs
- **Call Failure**: Higher call failures strongly correlated with churn (technical issues drive dissatisfaction).
- **Complains**: Complaints were a direct indicator of dissatisfaction.
- **Usage Patterns**: Low "Seconds of Use" or "Frequency of Use" signaled disengagement.
- **Customer Value**: Lower-value customers were more likely to churn.
- **Subscription Length**: Newer customers showed higher churn risk.

Next Steps & Recommendations

1. Model Improvements

- Address Class Imbalance: Use SMOTE or ADASYN to handle class imbalance and improve minority class prediction.
- Feature Engineering: Create interaction terms (e.g., "Call Failure per Usage") or temporal features.
- Advanced Models: Experiment with XGBoost, LightGBM, or neural networks for better performance.
    - Completed Experimentation with XGBoost and shared results with SHAP Interpretation. (Please refer Appendix section)

2. Deployment Strategies (for Customer Retention, Reactive >> Proactive approach)

Cisco Confidential

- Real-Time Integration: Embed the model into Customer Relationship Management (CRM) systems systems to flag high-risk customers during service calls.
- Reactive to Preemptive/Proactive approach: Leverage the prediction insights to drive Proactive/Preemptive "Next best" customer interactions to avoid churn rather than a reactive approach.
- Automated Alerts: Trigger automated retention offers and/or assign Customer Success Managers, Customer Success Specialists to key accounts/regions which have predicted churners. Leverage Email/Chat/Outbound call campaigns for proactive value delivery and customer intimacy, retention.

3. Ethical Considerations

- Bias Mitigation: Audit the model for fairness across demographics (e.g., age groups or regions).
- Transparency: Use SHAP/LIME to explain predictions to customers and build trust.
- Discuss and implement Responsible AI (RAI) framework to protect customer privacy data/assets.

4. A/B Testing

- Test retention strategies on a subset of high-risk customers and measure churn reduction compared to a controlled group.

5. Continuous Monitoring

- Retrain the model quarterly with fresh data to adapt to changing customer behavior.
- Retrain the model for seasonal data as well and for specific industries (for eg., for Healthcare customer buying SaaS subcriptions, do model training during Open Enrollment phase of the year, for retail customer during Thanksgiving, Christmas time et al)
- Track feature importance shifts over time (e.g., new pain points like "low adoption", "not signing up for new features", "customer stuck in a specific lifecycle" and not progressing to take best benefits of the subscription).

Final Recommendations:

- Prioritize campaigns to take care of "Basic subscription (base plan)" customers, have a dedicated customer success manager, proactive outbound campaign, discounts to uplift the customer from basic subscription to enhanced/premium subscriptions.
- Solve for technical issues (call failures) and improving customer service (reducing complaints) while deploying the model to target at-risk customers. This holistic approach will maximize retention and profitability.

Cisco Confidential

- Use XGBoost modeling to perform initial controlled environment/production trials, leverage SHAP to interpret best preforming model results, derive business insights and derive next best actions. Refer to additional experimentation and analysis of XGBoost outperformed and SHAP implementation provide key insights on model interpretation and performance as indicated in overall summary and plots.

Additional considerations may also include

- Hyperparameter tuning - Further experimentation with hyperparameter tuning for all models, particularly the Logistic Regression, to see if we can achieve better performance
- Interaction Terms: Based on EDA (especially the correlations and bar charts), create interaction terms between features that seem to have a combined effect on the target variable. SHAP interpretation with all models will help here.
- Polynomial Features: Consider adding polynomial features to capture non-linear relationships.
- Feature Selection/Dimensionality Reduction:
    - Regularization: For Logistic Regression, experiment with L1 (Lasso) or L2 (Ridge) regularization to reduce overfitting and potentially improve generalization.
    - PCA/Feature Importance: Use PCA or feature importance from a tree-based model (e.g., Random Forest) to select the most relevant features and reduce dimensionality.
- Recommend incorporating external data sources (e.g., consumer behvior, churn due to price vs competition offers) to enrich contact center feature set

Cisco Confidential

Appendix:

Experimentation with XGBoost and SHAP Interpretation

In addition to the above, performed Extreme Gradient Boosting classifier based analysis. Listed below are the XGBoost classifier before and after hyperparameter tuning. XGBoost model performed exceptionally well with high accuracy (97%), precision (91%), and an outstanding AUC score (0.99). It effectively identifies most churners while keeping false alarms low. However, addressing false negatives could further enhance its ability to detect all potential churners, making it even more robust for real-world applications like customer retention strategies.

1. Performance Metrics

• Accuracy: 0.97 (97%) The model correctly classifies 97% of all predictions, indicating high overall performance.

• Precision: 0.91 (91%) Among all instances predicted as "Churn," 91% are actual churners. This shows the model has a low false positive rate.

• Recall: 0.87 (87%) The model identifies 87% of actual churners, meaning it performs well in detecting churn but misses some cases (false negatives).

• F1 Score: 0.89 The F1 score balances precision and recall, showing the model is well-rounded in handling both false positives and false negatives.

• ROC AUC Score: 0.99 A near-perfect score of 0.99 indicates excellent separability between the "Churn" and "No Churn" classes.

2. Confusion Matrix

The confusion matrix provides a breakdown of predictions:

• True Negatives (523): The model correctly predicts "No Churn" for 523 customers.

• False Positives (8): Only 8 customers were incorrectly predicted as "Churn" when they did not churn.

• False Negatives (13): The model misses 13 actual churners, predicting them as "No Churn." • True Positives (86): The model correctly identifies 86 customers who actually churned. Insights:

• The model performs exceptionally well in predicting "No Churn" cases, with only a small number of false positives.

• While recall is strong at 87%, the false negatives (13) suggest there is room for improvement in capturing all churners.

Cisco Confidential

3. ROC Curve

The ROC curve evaluates the trade-off between the true positive rate (recall) and the false positive rate:

• The curve is very close to the top-left corner, indicating excellent performance.

• AUC = 0.99 confirms that the model has near-perfect discrimination between the two classes. Insights:

• The high AUC value reflects that the model is highly effective at distinguishing between churners and non-churners across different thresholds.

Key Strengths of XGBoost

1. High Accuracy: The model achieves an impressive accuracy of 97%, making it reliable for most predictions.
2. Strong Precision: With a precision of 91%, it minimizes false positives, which is crucial for avoiding unnecessary interventions for non-churning customers.
3. Excellent AUC: The high AUC score demonstrates that the model effectively separates churners from non-churners.

Areas for Improvement

1. False Negatives: While recall is strong, reducing the number of false negatives (13) would further improve the model's ability to capture all churners.
2. Possible actions: Adjusting decision thresholds or using techniques like oversampling/undersampling to handle class imbalance if present.
3. Class Imbalance Check: If churn cases are significantly fewer than non-churn cases, consider rebalancing the dataset to ensure better recall without sacrificing precision.

In addition to the above, performed Extreme Gradient Boosting classifier based analysis. Refer to the https://github.com/harishlv777/Capstone_CC_Churn/blob/main/plots/Capstone_CC_Churn_plots_v1.pdf to review XGBoost classifier performance before and after hyperparameter tuning.

1. Initial Model Performance

- Accuracy: 94.76% (indicates the overall correctness of predictions).
- AUC (Area Under the Curve): 0.9793 (shows high discrimination ability between classes).
- Precision, Recall, F1-Score:
    - Class 0 (majority class): High precision (0.96), recall (0.98), and F1-score (0.97).

Cisco Confidential

  - Class 1 (minority class): Lower precision (0.88), recall (0.77), and F1-score (0.82), indicating some difficulty in correctly identifying minority class instances.
- Confusion Matrix:
  - True Negatives (TN): 521
  - False Positives (FP): 10
  - False Negatives (FN): 23
  - True Positives (TP): 76
  - The model misclassifies some instances, especially for Class 1.

2. Tuned Model Performance

- After hyperparameter tuning with parameters like learning_rate=0.1, max_depth=7, n_estimators=200, and subsample=0.8:
- Accuracy: Improved to 96.83%.
- AUC: Increased to 0.9866, indicating better class separation.
- Precision, Recall, F1-Score: - Class 0: Precision, recall, and F1-score improved slightly. - Class 1: Significant improvement in precision (0.92), recall (0.87), and F1-score (0.90), reflecting better handling of the minority class.
- Confusion Matrix:
  - TN: 524
  - FP: 7
  - FN: 13
  - TP: 86
  - The number of misclassifications decreased for both classes.

XGBoost Key Takeaways

- The XGBoost model performs well initially but shows bias toward the majority class.
- Hyperparameter tuning significantly improves the model's performance, particularly for the minority class, as seen in better precision, recall, and fewer misclassifications.
- The tuned model is more balanced and effective at distinguishing between the two classes while maintaining high overall accuracy and AUC.

SHAP Interpretation

SHAP based interpretation is performed to gain deeper insights into the decision-making process of XGBoost model, helping validate its reliability and interpretability. The SHAP plot reveals that customer usage metrics (Status, Frequency_of_Use) and service quality indicators (Call Failure, Complains) are the most influential features in the model's predictions. These insights can guide targeted interventions, such as improving service reliability or addressing complaints to enhance predictive accuracy and customer satisfaction.

Cisco Confidential

This SHAP (SHapley Additive Explanations) summary plot https://github.com/harishlv777/Capstone_CC_Churn/blob/main/plots/Capstone_CC_Churn_plots_v1.pdf shows how each feature impacts the model's predictions.

- x-axis (Impact on Model Output)
- Negative values decrease the likelihood of churn.
- Positive values increase the likelihood of churn.
- y-axis (Model Features)
- The order of the features on the Y-axis is based on overall importance.
- Color (Feature Value Intensity)
- Blue indicates lower feature values.
- Red indicates higher feature values.

SHAP Key Insights

- Features at the top of the plot have the highest impact on the model's predictions, while those at the bottom contribute less.
- Top Impactful features
  - Status: This feature has the most significant influence, with a wide range of SHAP values indicating strong predictive power.
  - Frequency_of_Use and Seconds_of_Use: These usage-related metrics also play critical roles in determining predictions.
  - Call Failure and Usage_per_Sub_Day: Indicators of service quality and usage patterns are highly relevant.
- Lesser Impactful features
  - Features like Age_30 and Callback_Count, located near the bottom, have minimal influence on predictions. These might be less relevant for decision-making or could be candidates for removal during feature selection.
- In Status, higher values (red) are associated with positive impacts on predictions, while lower values (blue) contribute negatively.
- In Complains, higher values (red) negatively impact predictions, suggesting that more complaints correlate with unfavorable outcomes.
- Wider distributions of SHAP values for features like Status and Frequency_of_Use indicate variability in their influence across different samples.
- Narrower distributions for features like Age_Group_Numeric suggest consistent but less impactful contributions.
- Features related to customer usage patterns (Frequency_of_Use, Seconds_of_Use) are critical for understanding customer behavior.
- Service quality metrics (Call Failure, Complains) are key drivers for predicting outcomes, emphasizing their importance in improving customer satisfaction.

Cisco Confidential

Plots

https://github.com/harishlv777/Capstone_CC_Churn/blob/main/plots/Capstone_CC_Churn_plots_v1.pdf

Files

- SiddiCC_Churn.ipynb - Jupyter notebook
- data/SiddiCC_Churn_data.csv - Contact Center dataset
- plots/Capstone_CC_Churn_plots.pdf - plots supporting the analysis
- readings - crisp-dm-overview.pdf CRISP-DM methodology document
- readme.md - current file

Requirements

- Python 3.x, pandas, numby, matplotlib, seaborn, scikit-learn Note
- plot_helpers is required to render_plot
- Run pip list | grep plot_helpers to check if plot_helpers exists. If missing, either install it or replace render_plot with Matplotlib/Seaborn functions
- SHAP library for model interpretation

How to execute

- Clone the repository
- Build the environment with required packages, followed by Jupyter notebook execution.

Cisco Confidential

Appendix - Non-technical Report

Executive Summary

Prediction of Churn in SaaS business using Contact Center data and Machine Learning models.

"In today's subscription-based economy, high customer churn poses a significant threat to enterprises and service providers. This capstone project tackles this critical challenge by leveraging machine learning to predict customer churn risk within a contact center. By analyzing data encompassing customer interactions, agent performance, and demographic factors, the project aims to develop a predictive model capable of identifying customers at high risk of churn. This model will empower contact centers to proactively implement targeted retention strategies, ultimately boosting customer satisfaction and minimizing revenue attrition.

Problem Statement: Why do Customers leave?

Key factors contributing to churn were identified:

- **Basic Subscription Plans** – Customers on lower-tier plans tend to leave faster than premium users.
- **Frequent Complaints** – Customers who report multiple issues are more likely to leave.
- **Low Usage** – Less engagement with the service indicates a risk of churn.
- **Short Subscription Periods** – Customers with shorter subscriptions are more likely to cancel.

Solution

As part of this Capstone project a **machine learning model** is developed to **identify at-risk customers** before they churn. This allows businesses to proactively implement targeted retention strategies.

Key highlights include:

- **95% Accuracy** – The model successfully predicts potential churners.
- **Key Predictors** – Call duration, complaints, and subscription type were the top indicators.
- **Churn Indicators**: Customer complaints, Low usage of services, Short subscription lengths, Basic subscription plans
- **Proactive Strategies** – By identifying high-risk customers, businesses can take action **before** they leave.

Key Findings

1. **Customer Behavior**:
    1. Customers with basic subscriptions have a higher churn rate

Cisco Confidential

    2. Higher-value customers are less likely to churn
2. **Predictive Model**:
    1. Random Forest model achieved 95% accuracy in predicting churn
    2. Key predictors: Call duration, complaint history, and subscription type
3. **XGBoost Advanced model and experimentation** showed promising results. XGBoost model performed exceptionally well with high accuracy (97%), precision (91%), and an outstanding AUC score (0.99). It effectively identifies most churners while keeping false alarms low.
4. **SHAP interpretation demonstrated deeper insights**. The SHAP plot reveals that customer usage metrics (Status, Frequency_of_Use) and service quality indicators (Call Failure, Complains) are the most influential features in the model's predictions
5. Use XGBoost modeling to perform initial controlled environment/production trials, leverage SHAP to interpret best preforming model results, derive business insights and derive next best actions

Top Features Impacting Churn

- **Status**: Customers with Status = 1 (basic software subscription) have a higher churn rate compared to customers with Status = 0. It would be good to target the customers with basic subscription plans through campaigns, dedicated customer success managers, discounts, trial to uplift them to premium subscription with value added features. There more customers with “Basic Subscription (base plan)” who would cancel a subscription service than engaged customers with added features who actively use subscription programs
- **Call Failure**: Higher call failures strongly correlated with churn (technical issues drive dissatisfaction).
- **Complains**: Complaints were a direct indicator of dissatisfaction.
- **Usage Patterns**: Low "Seconds of Use" or "Frequency of Use" signaled disengagement.
- **Customer Value**: Lower-value customers were more likely to churn.
- **Subscription Length**: Newer customers showed higher churn risk

Recommendations

1. **Upgrade Basic Subscriptions**: Target customers with basic plans through:
    1. Personalized campaigns
    2. Dedicated customer success managers
    3. Discounts or trials for premium features
2. **Improve Technical Quality:** Reduce call failures to decrease dissatisfaction
3. **Proactive Outreach**: Use the predictive model to identify at-risk customers and intervene early
4. **Enhance Customer Value**: Develop strategies to increase usage and subscription length
5. **Streamline Complaint Resolution:** Implement efficient processes to address and resolve customer complaints quickly

Cisco Confidential

Implementation

1. Integrate the predictive model into the contact center's Customer Relationship Management (CRM) system
2. Train customer service representatives to use model insights for personalized interactions
3. Develop automated triggers for proactive retention strategies based on churn risk predictions
4. Regularly update and refine the model with new data to maintain accuracy

Expected Outcomes

- Reduction in overall churn rate
- Increased customer retention and lifetime value
- Improved operational efficiency in the contact center
- Enhanced customer satisfaction and loyalty

By combining **Model performance data and insights, AI/ML customer insights, and proactive outreach**, businesses can **retain more customers**, **improve loyalty**, and **increase profitability**. This ML based **data-driven approach** ensures that businesses stay ahead of churn before it becomes a problem.

Cisco Confidential