PLACEHOLDER FOR WEBSITE

LinkedIn

PLACEHOLDER FOR GITHUB

Olga Waldkirch, PhD

Identifying drivers of customer churn and designing targeted recommendations to improve retention.

February 2025

Table of Contents

Summary

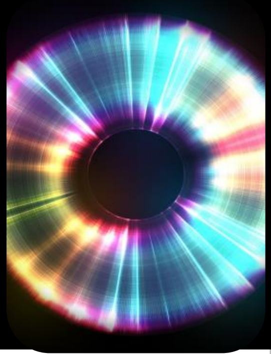
We have lost 26.5% of our customer base in the last six years. It has negatively impacted revenue, marketing spend, and our ability to re-invest in upgrades to keep up with competitors. To propose interventions for retention improvement, we developed a predictive model to identify at-risk accounts and contributing factors. In the process, we checked the impact of the customer service changes, implemented in 2021.

As a result, we propose a cross-functional task force to validate and expand initial findings and to oversee pilots of prioritized interventions, among which:

1. Re-evaluation of automatic enrollment into Technical Support or other incentives for this service.
2. Re-evaluation of incentives for longer-term contracts or other initiatives to promote such contracts.
3. Consideration of incentives for accounts with dependents.

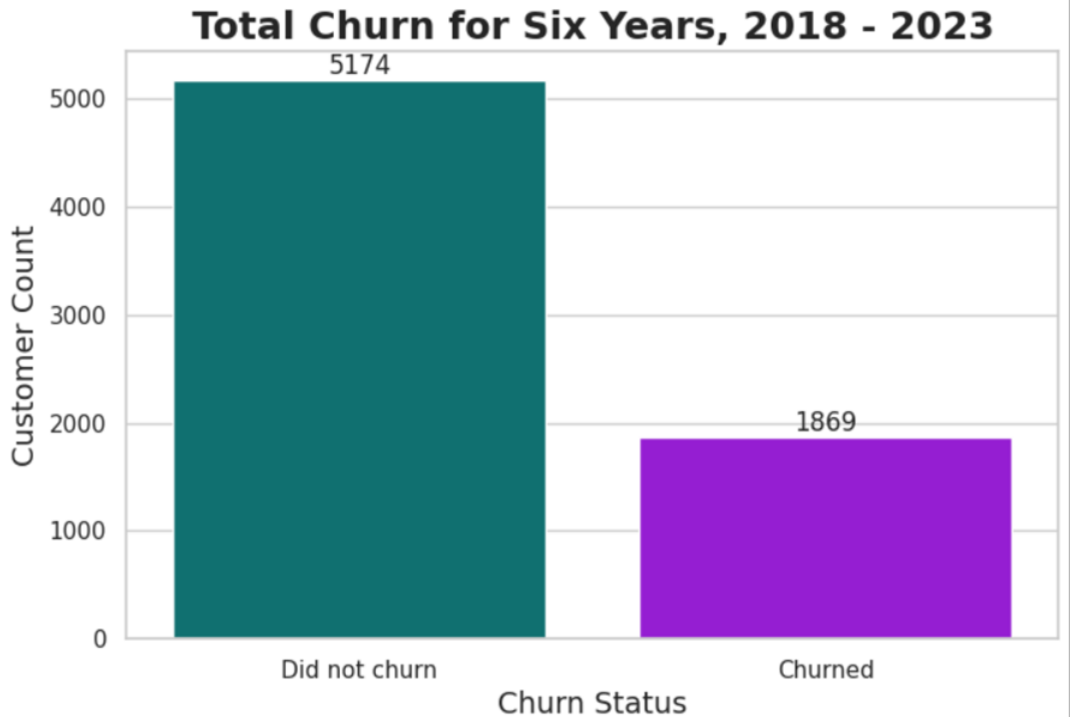
Problem Statement

Current state of customer churn



In six years, we have 1,869 customers leave the company, or

26.5% of the customer base lost



Insufficient data collection

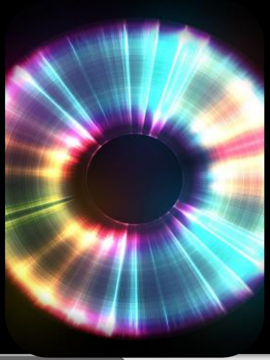
We capture reasons to leave for only 27% of churned customers. Thus, we do not know what motivated as many as 73% to cancel our services.

27% customer feedback captured

Top captured reasons for leaving

- Competitor offered better services (31%)
- Product dissatisfaction (19%)
- Attitude of support person (17%)

Customer services changes (2021)



Technical Support

Discontinued automatic enrollment in Technical Support as customers did not find value in automatic recurring fees for this service.

Incentives for Two-Year Contract

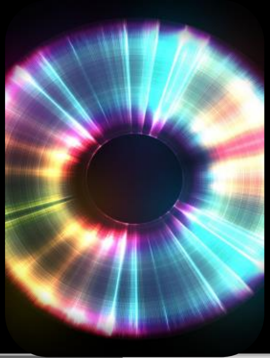
Discontinued incentives for two-year contracts as no significant difference was found between two-year, one-year, and monthly contracts.

Reduced Senior Citizen Discount

Current Reduced Senior Citizen discount by 10% in an effort to increase subscription revenue. It affected 14% of the customer base.

We need to understand what impact these changes have had on the churn rate.

Risks of sustained high churn rate



- Declining subscription and service revenue
- Reductions in marketing spend
- Delayed upgrades to infrastructure and services

Current Customer Lifetime Value to Customer Acquisition Cost Ratio

is below the industry standard and negatively affects our credit rating.

2.1:1

Call to Action

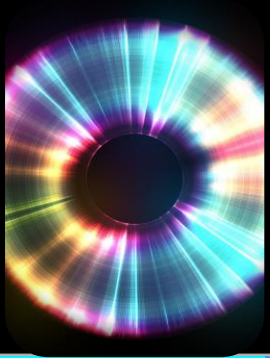
Understand the impact of the 2021 customer service changes on the churn rate and identify other top contributing factors. Design at-risk profiles to implement targeted initiatives.

Solution

Develop a classification model using the CRM data set of 7,043 customer records.

Classification Model

Data set



The original data set has 7,043 customer records over six years. As many as 1,869 have left our company, or 26.5%.

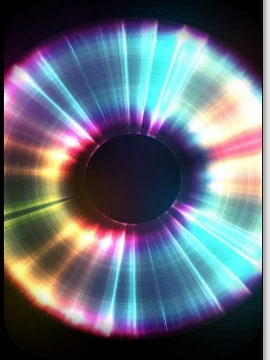
1,869 customers have canceled services

What is recorded for each customer

Top reasons shared

- Demographics
- Subscription services
- Contract type
- Payment method
- Monthly and total charges
- Status (churn or did not churn)

Please see Appendix for a full list.



Data preparation: cleaning, outliers, and missing values

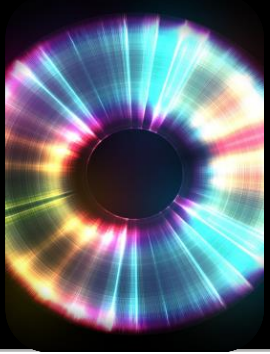
11 Customers with zero Tenure Months were assigned $0 for Total Charges.

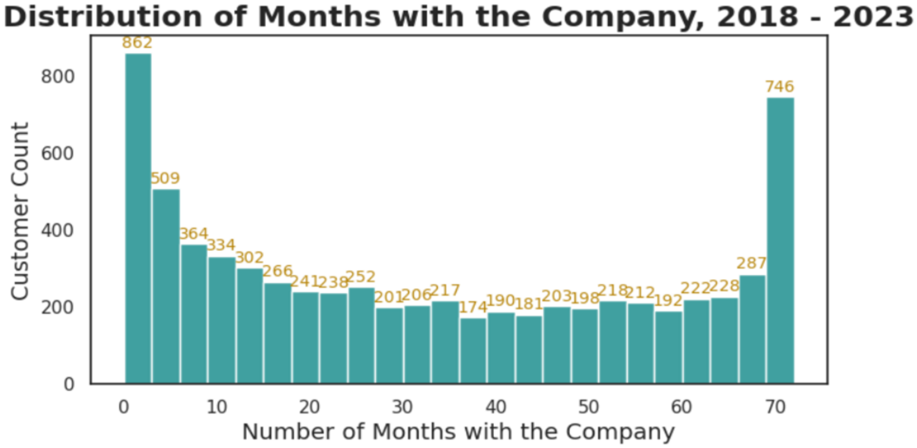
Columns with no predictive power were removed.

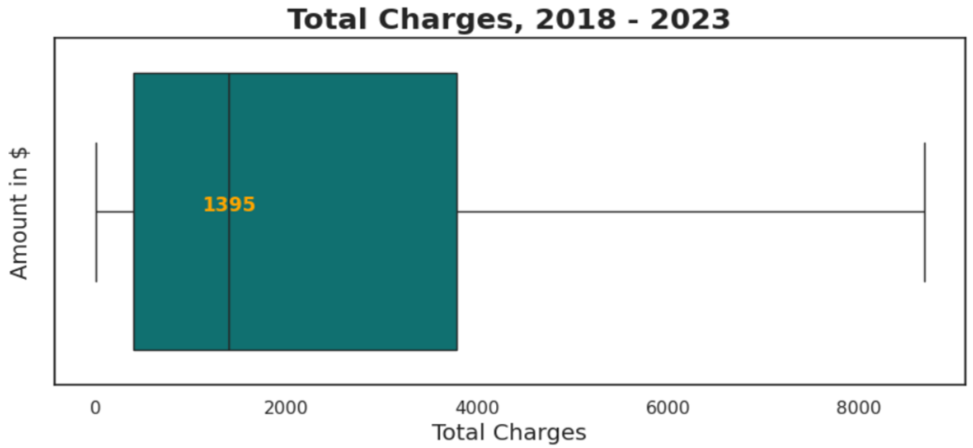
Categorical variables were encoded for the use in the algorithm.

No outliers were detected.

Initial findings: customer tenure and total charges



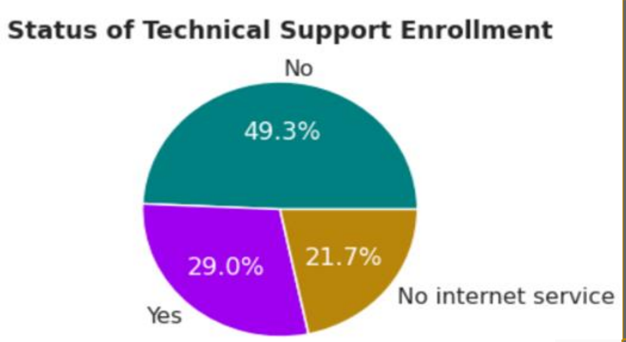




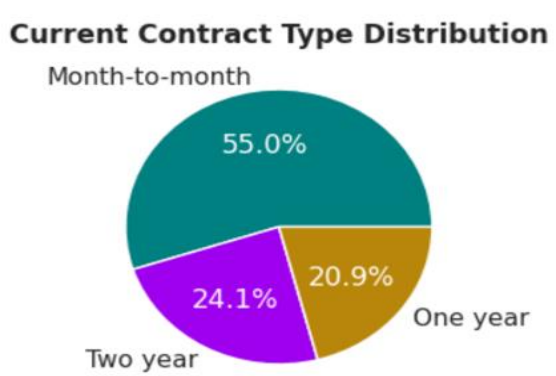
As many as 1,045 customers stayed with the company beyond five years. However, we want to see more customers shifting to longer tenure, i.e., improve retention.

Consistent with the month tenure, we see the majority of the total charges over the median of $1,395. Longer tenures result in higher revenues, positively impacting customer lifetime value.

Initial findings: status of customer service changes of 2021



Almost half of our customers are not enrolled into Technical Support, with as many as 22% choosing not to sign up for the Internet service altogether. As one of the changes made in 2021, this deserved further researching.



Over half of our customers currently hold the month-to-month contract while the one-year contract is the least popular option. **About a quarter have the two-year contract, incentives for which were discontinued in 2021.** The relationship between churn and contract type will also be evaluated.

Initial findings:
what can we expect to influence churn?

We analyzed over 7,000 accounts to identify leading drivers of churn.

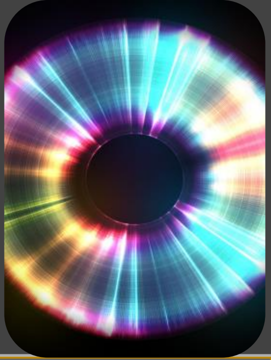
As the first step, we aimed to see how service cancellation is correlated with what we know about our customers: demographics, contract type, subscriptions, and tenure with the company.

Possible contributing factors

- Contract Type (month-to-month, one-year, two-year)
- Tenure months (number of months with company)
- Dependents (Y/N)
- Device protection (Y/N)
- Technical support (Y/N)
- Payment method (e-check, credit card, mail)
- Paperless billing (Y/N)
- Streaming movies (Y/N)
- Streaming TV (Y/N)
- Online back-up (Y/N)
- Online protection (Y/N)
- Online security (Y/N)
- Internet Service (Y/N)

Please see Appendix for correlation matrices and p-values.

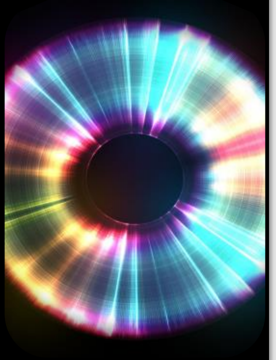
Further analysis
to qualify potential drivers of churn



The exploratory analysis reveals many possible influencing factors but does not show direction of the relationship or narrow the list down to the most impactful drivers.

We will apply further predictive modeling to qualify potential drivers and magnitude of their impact on churn.

These findings will aid us in identifying at-risk customer profiles and designing targeted interventions to improve retention.



Choice of algorithm: XGBoost with GridSearchCV (cross-validation)

Well-equipped to handle imbalanced data.

Uses early stopping which prevents overfitting to the majority class (in our case, the “non-churn” class).

Captures complex relationships, even in imbalanced data.

Can be combined with GridSearch and Cross Validation to find optimal parameter combination.

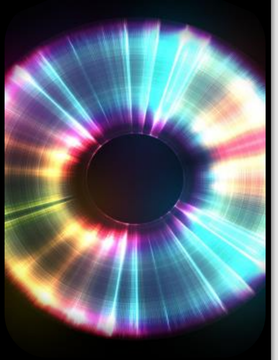
Adding interpretability and transparency: SHAP (SHapley Additive exPlanations)

Enhances the XGBoost model by providing transparency into each feature’s contribution to prediction.

Provides both global and local explanations, i.e., explanations for individual predictions can be displayed.

Shows which feature is most important in making predictions (global behavior of the model).

Shows which features increase or decrease the likelihood of churn.



Modeling procedure

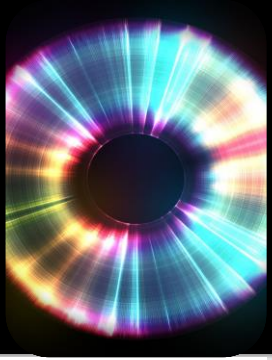
The data set was split into train (5,282 records or 75%) and test (1,761 records or 25%) subsets, each representing 26.5% of the churn class.

A preliminary model was fit and assessed for performance.

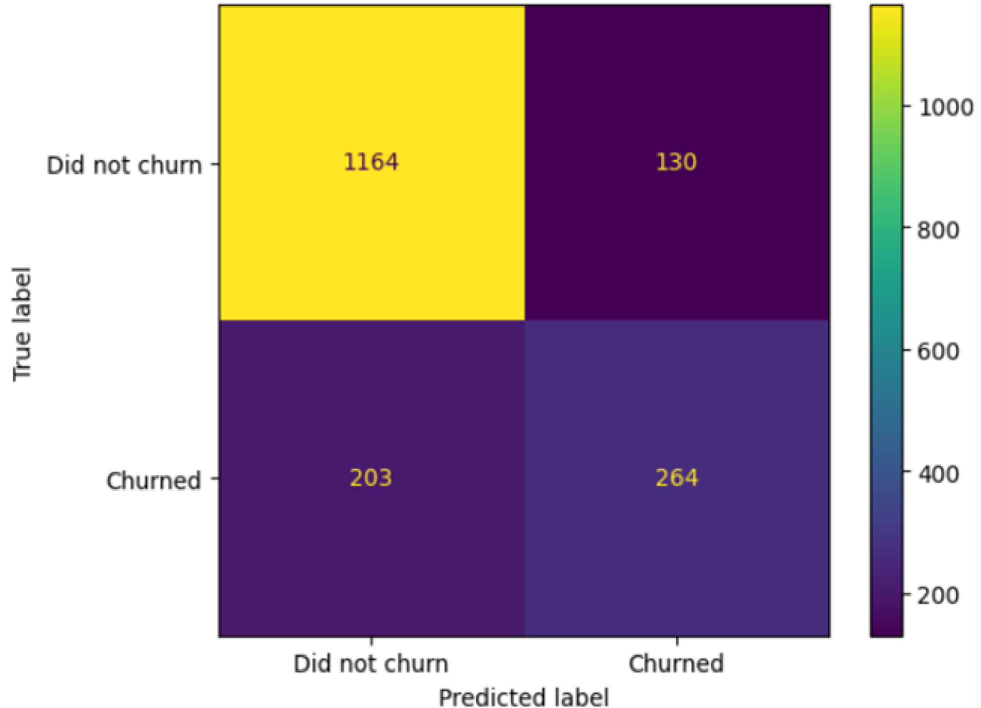
The preliminary model was improved by hyperparameter tuning and scaling of the minority class (churn).

The final model was assessed for performance.

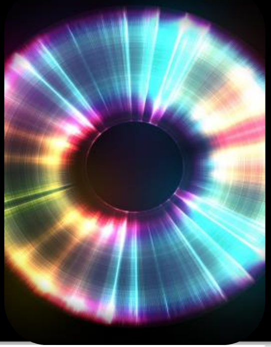
Final model performance: churn classification and trade-offs



We pursued the reduction of False Negatives, i.e., missed churned customer records, in improving the model. In the final version, the rate of missed churn was reduced from 43% to 17%% and the rate of identifying actual churn was improved from 57% to 83%, but these gains came with a trade-off in the increase of False Positives, i.e., false alarms from 10% to 27% and reduction in identifying non-churn customers from 90% to 73%.

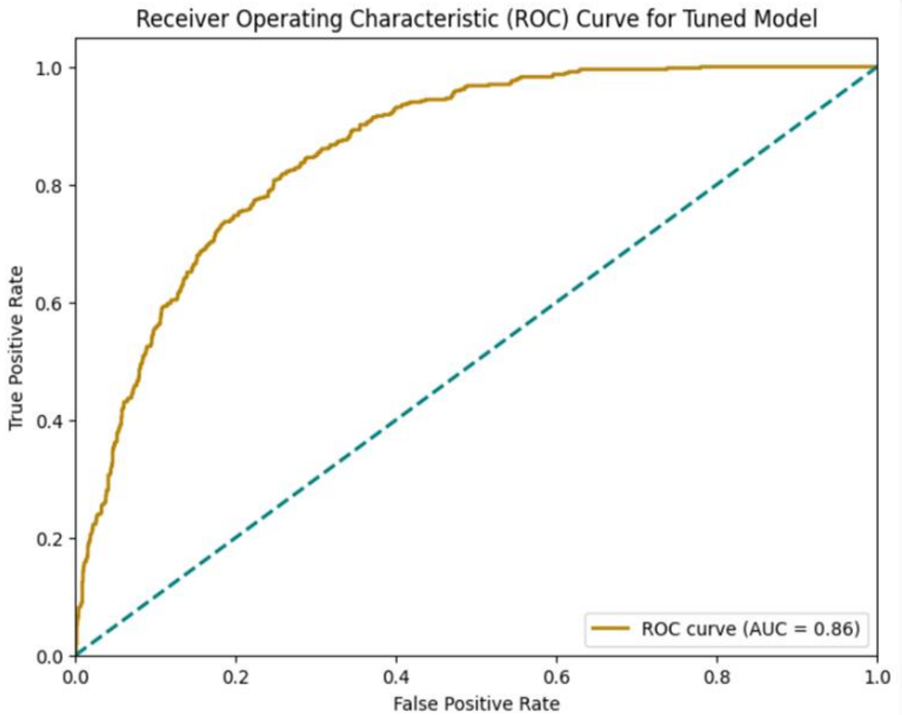


Final model performance: how well does it predict?

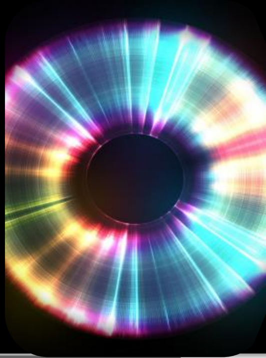


Under the final model, we have an 86% chance of each churned account classified correctly. In addition, it correctly classifies 75% of all the records and demonstrates the following performance metrics:

Accuracy Score (% of all correctly classified records, churm and non-churn)	75.35%
Precision Score (% of correctly classified churned records out of all *predicted* churned records)	52%
Recall Score (% of correctly classified churned records out of all *actual* churned records)	82%
F-1 Score (how well does the model balance between missed churn and false alarms, from 0 to 100?)	0.64



What the model does well and where it needs further improvements



Recalling churned records

Out of all actual churned cases, the model correctly recalls over 80% of such records.

Capturing easy-to-miss cases

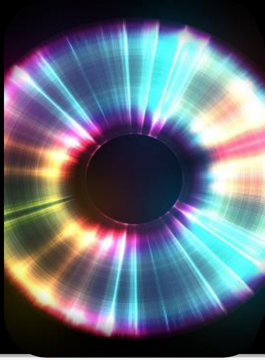
The rate of missed churned cases is 17% which is an acceptable metric for highly imbalanced data.

Producing false alarms

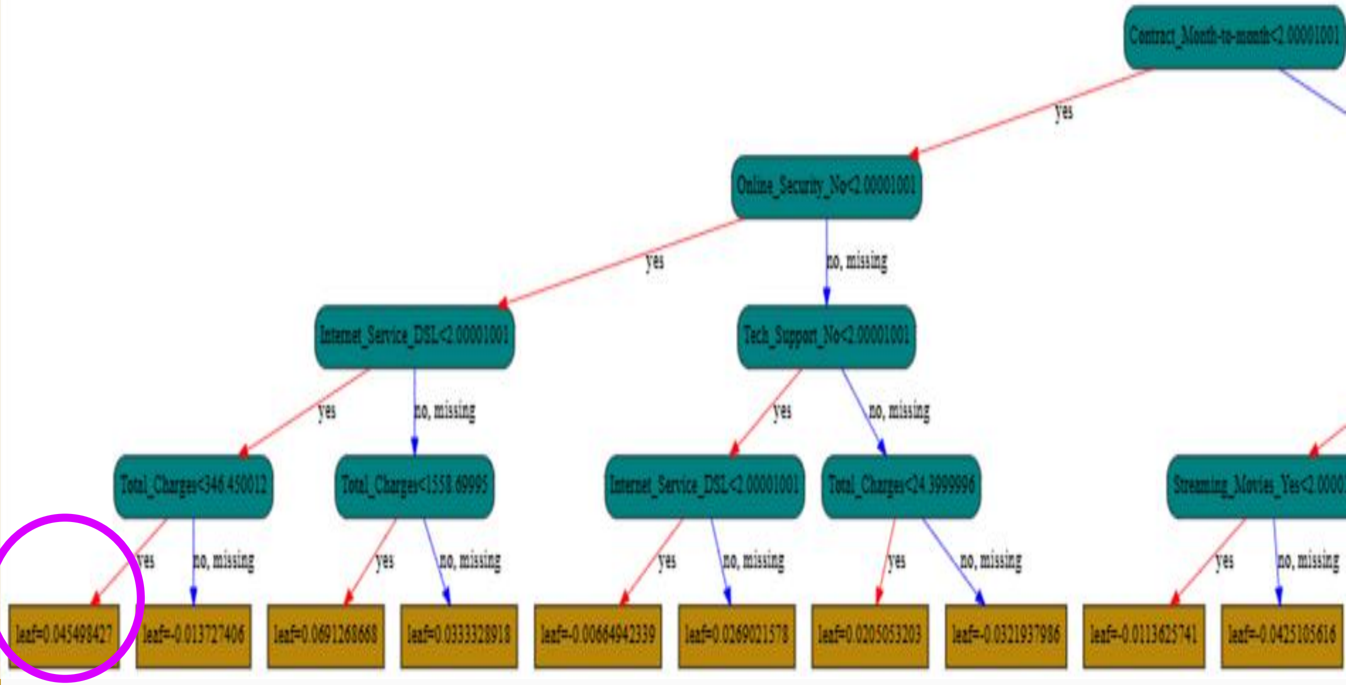
Extra emphasis on the churn class creates a side effect of an increase in false alarms, customers predicted to churn but who actually stayed with the company.

We will continue working on tuning the model in accordance with this task force’s recommendation.

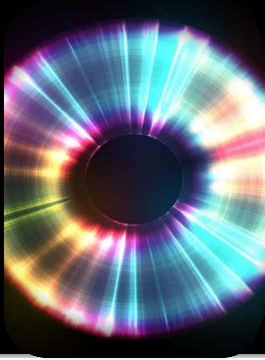
What can the model tell us about an individual customer?

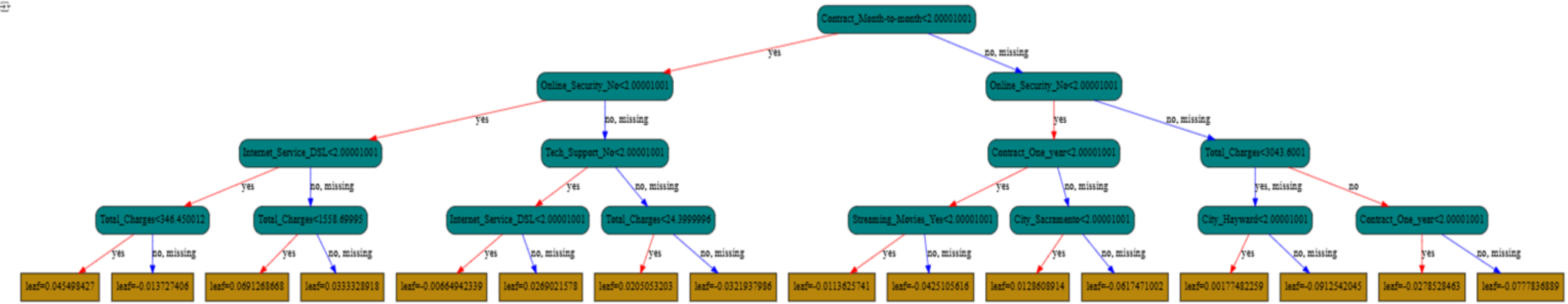


Taking the left-most classification as an example: **a customer with a contract other than month-to-month, enrolled into Online Security option, with Total Charges of less than $3,000, and residing in a city other than Hayward, is not likely to churn** (probability of churn is 45%, i.e., less than 50%).



What else can the decision tree tell us?



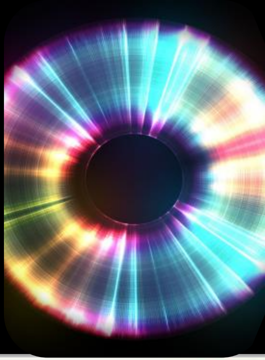


The most influential feature of distinguishing between churn and non-churn customers is between the contract type month-to month and the other two types of contract, one-year and two-tear.

Group A, with one-year and two-year contracts, is then split by enrollment into online security, followed by the total charges above or under $3,000 and again by the type of contract, now separating the one-year type from the two-year type.

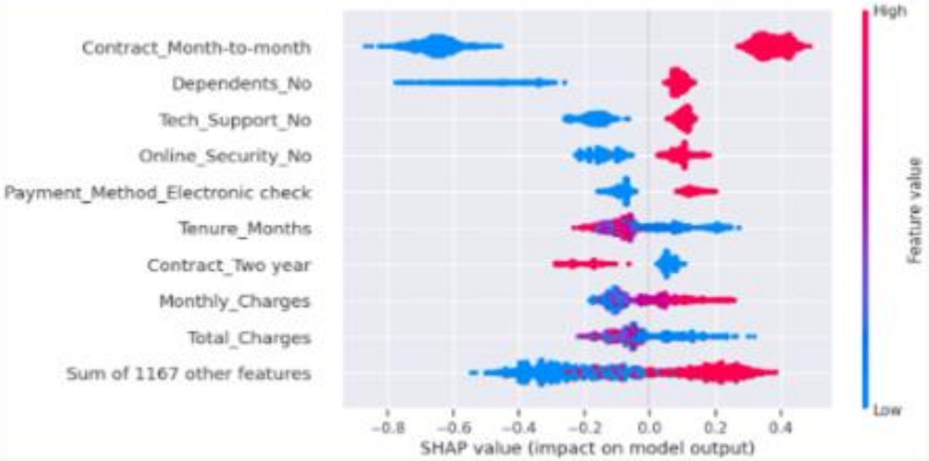
Group B, with month-to-month contracts or missing records, is also first split on enrollment into online security but then followed by enrollment into technical support and Internet services.

Adding the feature importance analysis for transparency and explainability

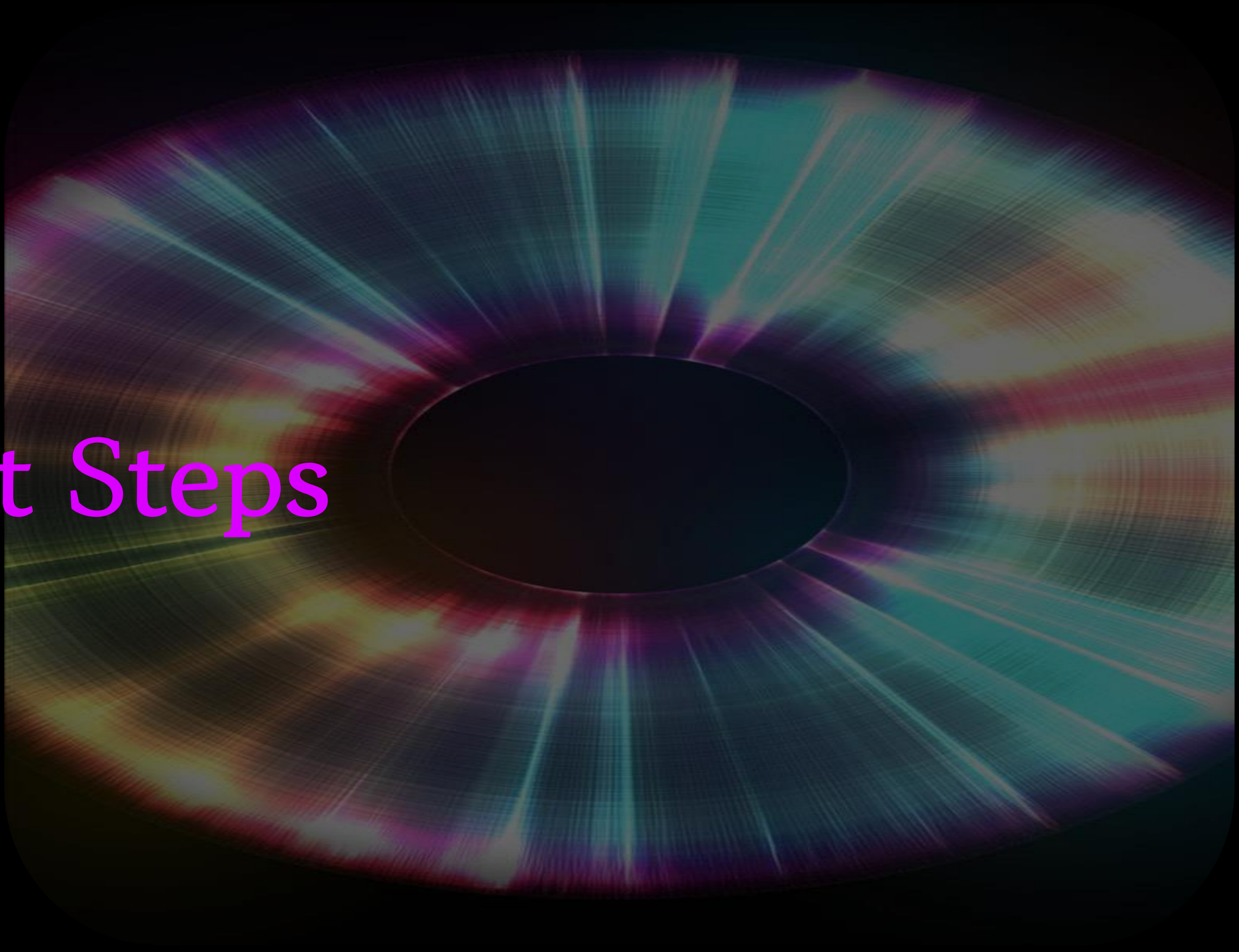


This ranked list of the most influential features shows a few insights:

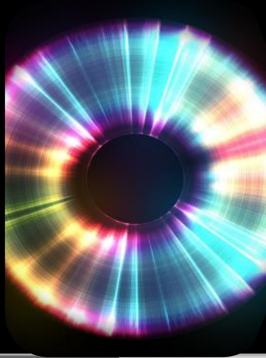
- The most important feature in distinguishing between churn and non-churn is whether the customer holds the month-to-month contract or the other two types.
- Customers with dependents are less likely to cancel services.
- Enrollment in online security and technical support strengthens retention.
- There are several cases when customers with higher tenure months still left our company.



Next Steps



Re-visiting customer services changes (2021)



Technical Support

Customers without the technical support feature are more likely to cancel services. This factor, however, is the third most influential feature, according to the model.

Incentives for Two-Year Contract

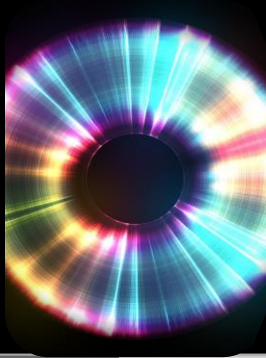
Customers with month-to-month contracts are more likely to cancel services. In fact, this type of contract is the most influential factor. Currently, month-to-month contracts represent 55% of accounts.

Reduced Senior Citizen Discount

At this stage, it appears that the reduction did not have a traceable effect as the analysis does not show the Senior Citizen status as a contributing factor to churn.

We propose re-evaluating the changes to automatic enrollment in technical support and incentives for longer contracts.

Proposed targeted interventions



Incentives for longer contracts

Discontinued automatic enrollment in Technical Support as customers did not find value in automatic recurring fees for this service.

Incentives for dependents

Discontinued incentives for two-year contracts as no significant difference was found between two-year, one-year, and monthly contracts.

Automatic Enrollment Consideration

Current Reduced Senior Citizen discount by 10% in an effort to increase subscription revenue. It affected 14% of the customer base.

These are initial recommendations,

pending further analysis of at-risk accounts.

February 2025
Launch a task force to complete at-risk profiles and design interventions

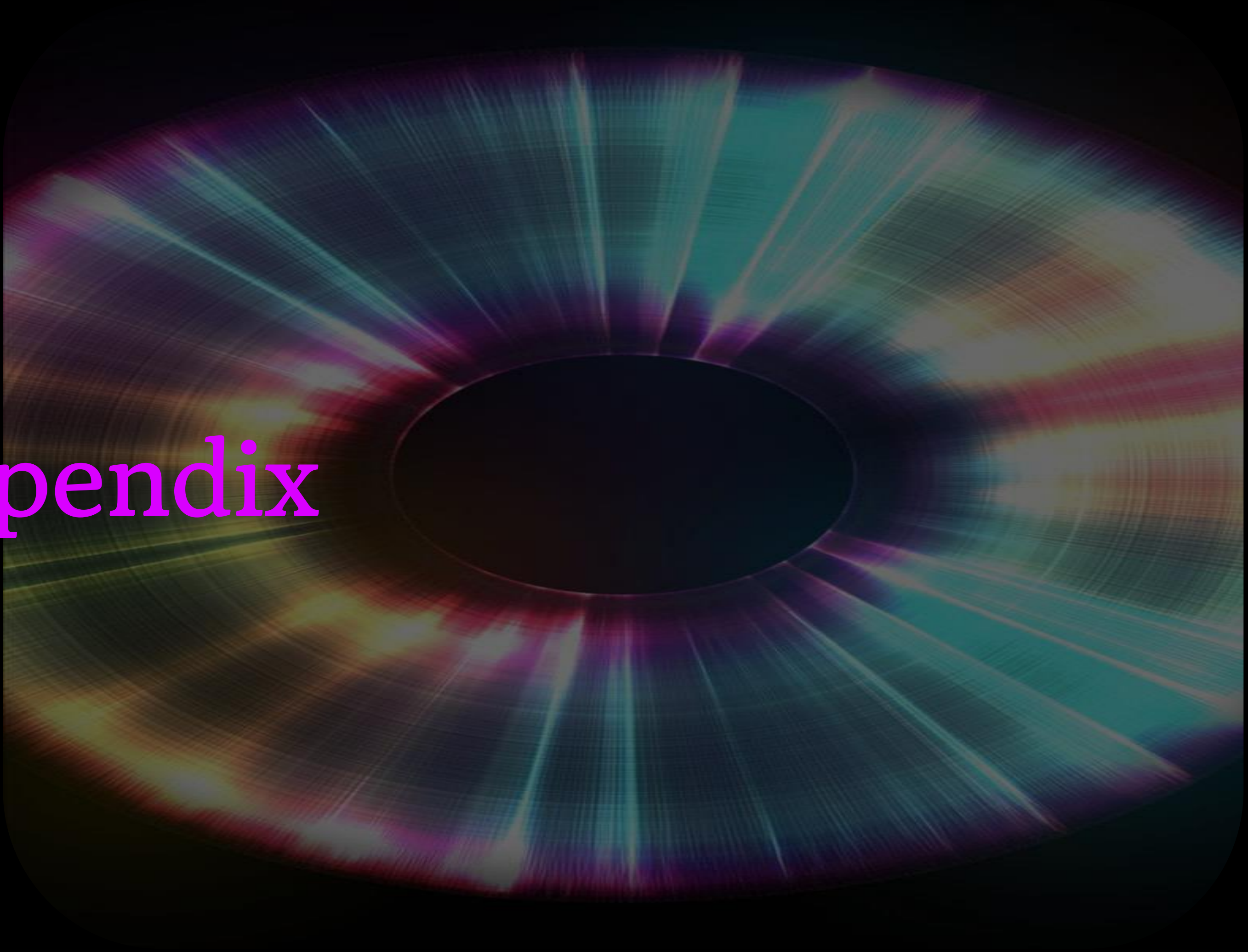
April 2025
Approve and fund prioritized interventions, establish KPIs

May 2025
Implement interventions in select markets as a pilot

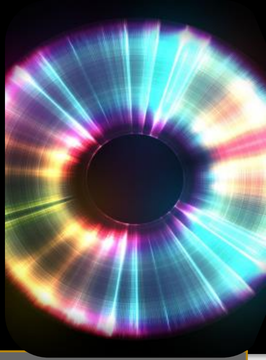
October 2025
Assess the pilot, make adjustments, decide on extension

May 2026
Assess lagged KPIs of the pilot, make adjustments

Appendix

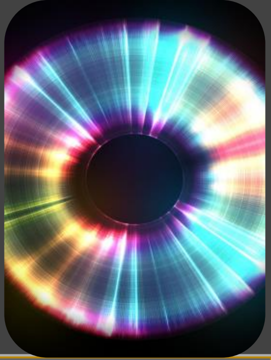


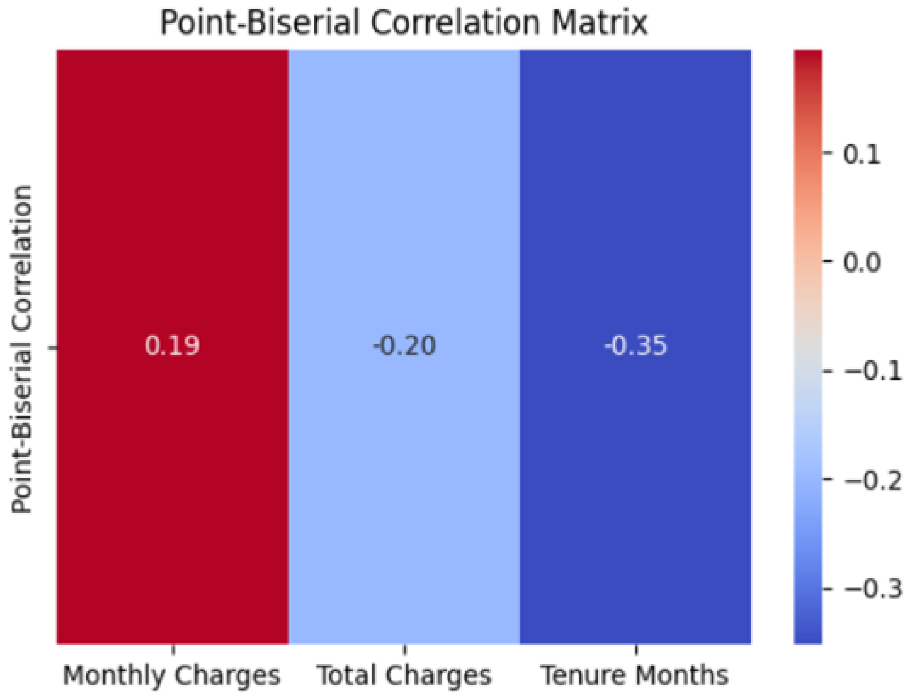
Appendix I: full list of data types, collected for each customer



Customer ID	Dependents	Online Backup	Paperless Billing	State
City	Tenure Months	Device Protection	Payment Method	Country
Zip Code	Phone Services	Tech Support	Monthly Charges	Churn Label
Gender	Multiple Lines	Streaming TV	Total Charges	Churn Reason
Senior Citizen	Internet Service	Streaming Movies	Latitude	
Partner	Online Security	Contract Type	Longitude	

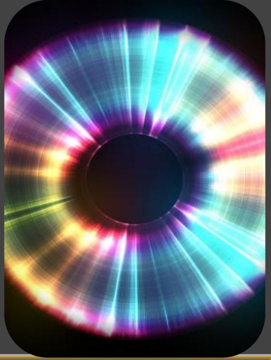
Appendix II: point biserial correlation for numeric features



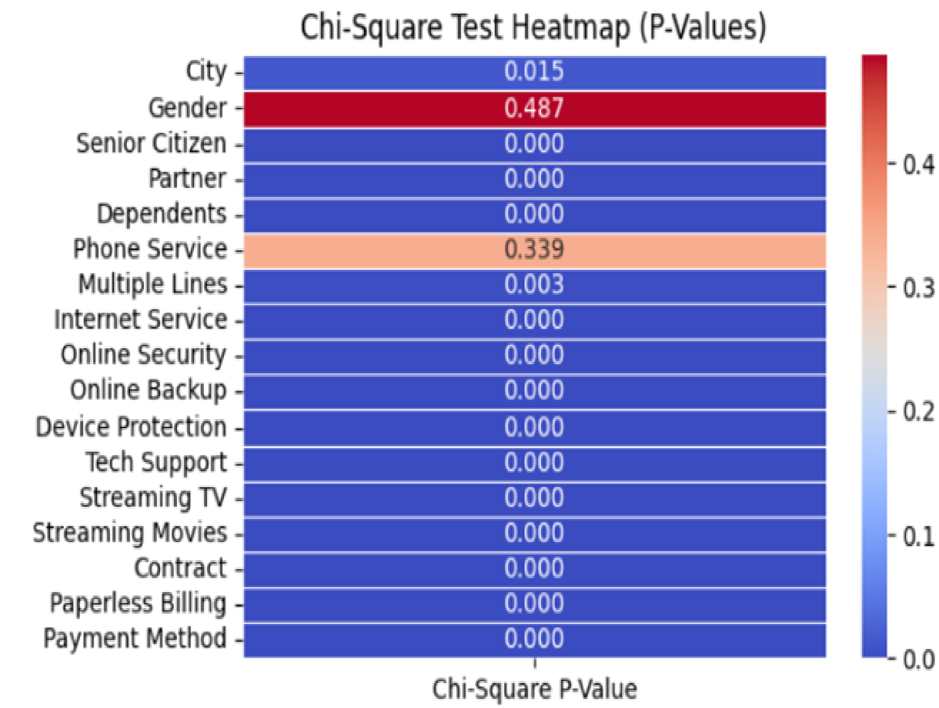


In the case of a binary target variable such as churn (churn = 1, did not churn = 0) and numeric continuous predictors, an adapted version of Pearson's correlation, Point Biserial Correlation, is recommended. Here we see

- weak positive correlation between Monthly Charges and churn, i.e., as monthly charges increase, so does the likelihood of churn.
- weak negative correlation between Total Charges and churn, i..e, as the amount of total charges goes up (longer tenure with company), the likelihood of churn reduces.
- low moderate negative correlation between churn and Tenure Months, i.e., longer tenures are associated with the reduction in the likelihood of churn.

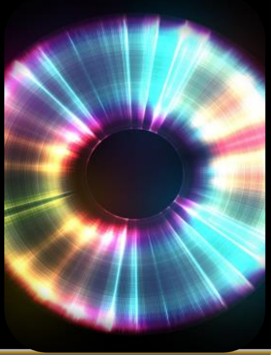


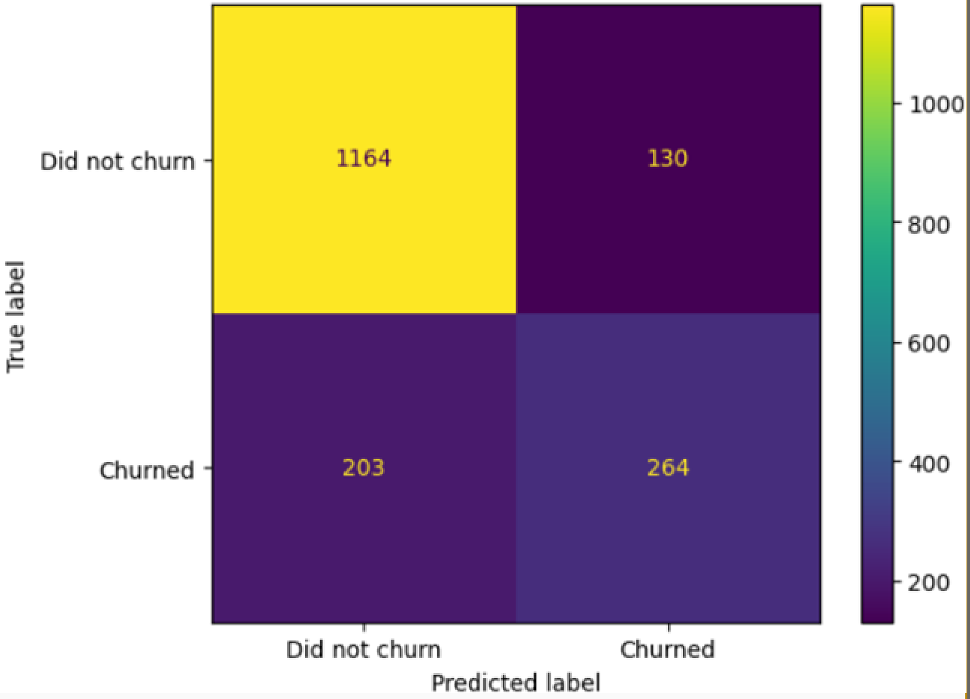
Appendix III: Chi-squared independence test



In the case of a binary target variable such as churn (churn = 1, did not churn = 0) and categorical predictors, the Chi-squared independence test is recommended. This table shows the p-values of the relationship between churn and the categorical variables where the values of $p \leq 0.05$ signify a statistically significant relationship, albeit without direction, which requires further analysis such as Logistic Regression or SHAP.

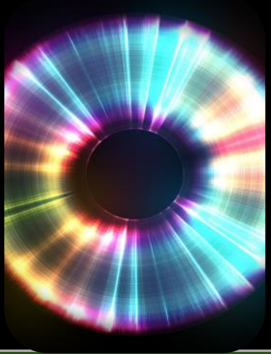
Appendix IV: Preliminary Model Performance Confusion Matrix

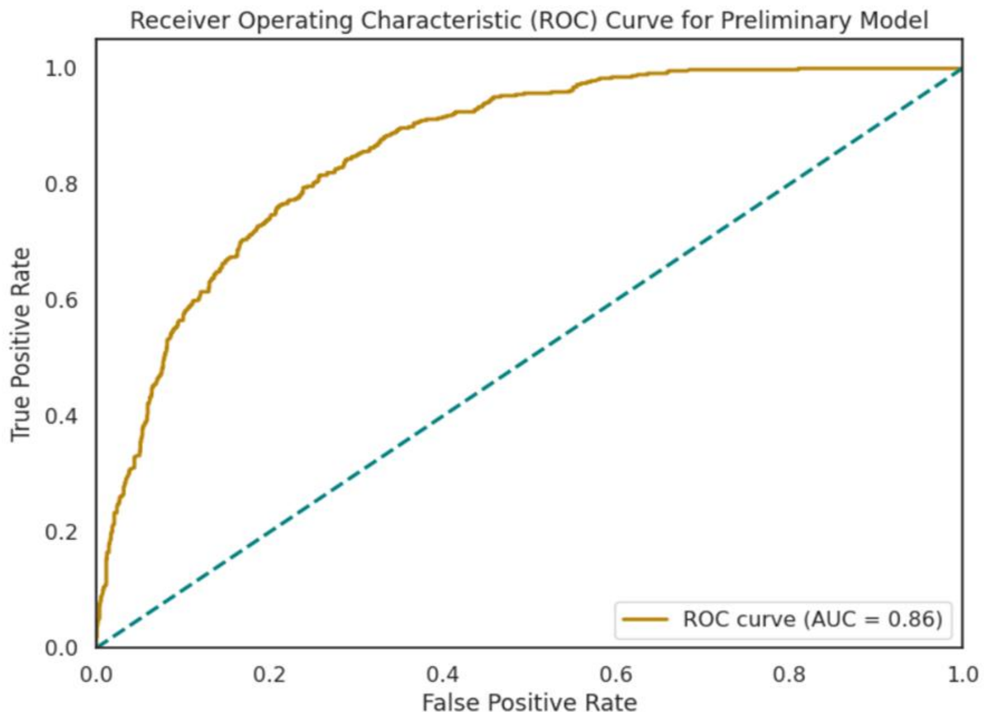




As expected, the imbalance of classes affects the model performance. As much as 43% (203) of churned records are misclassified. However, the model performs well predicting True Negative records, i.e., customers who stayed with the company.

Appendix V: Preliminary Model Performance Receiver Operating Characteristic and Metrics

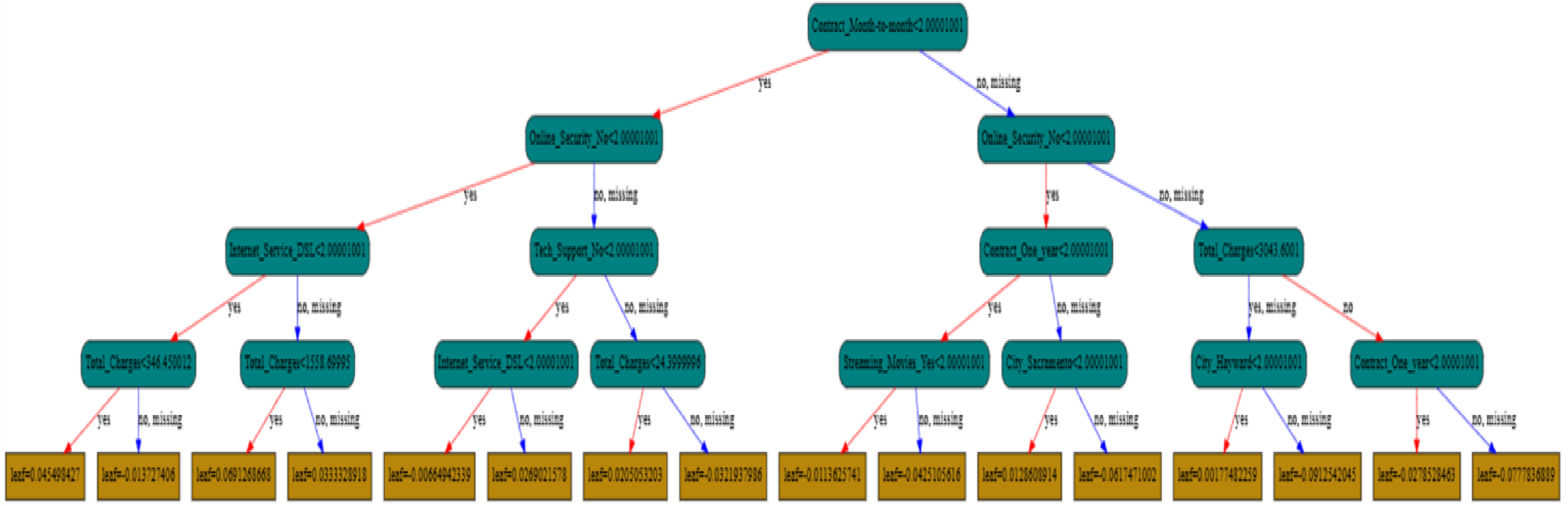




The Area under the Curve metric of the preliminary model is 86%, i.e., if we select a random churned customer, we have an 86% chance that it will be classified correctly. In addition, it shows the following performance metrics:

Accuracy Score	81.09%
Precision Score	67.01%
Recall Score	56.53%
F-1 Score	0.61

Appendix VI: Full Decision-Tree



Appendix VII: SHapley Additive exPlanations (global)

For example, the red points for monthly charges are mostly on the right side indicating that higher monthly payments drive churn; the lower values for monthly charges, which are blue points, are on the left, thus decreasing the likelihood of churn. There are some red points on the left side too, indicating the customers with higher monthly bills could still keep their services. Their individual profiles may reveal other contributing factors.

The features are ranked in the descending order of importance.

1. First, distinguish between red and blue points for a feature: red corresponds to higher values and blue to lower values of the feature. For categorical features, the higher value is 1 (feature present) and lower value is 0 (feature not present)
2. Second, look on which side of the graph points appear: if they appear on the right, i.e., show positive SHAP value, then they contribute to churn; if they appear on the left, i.e., show negative SHAP values, they deter churn.