**Final Paper**

Kyle Converse

Franklin University

DATA 695: Capstone

May 10, 2025

**Introduction**

The purpose of this project is to develop a robust, precise, actionable model to predict customer churn for Telco, a large telecommunications company. The project will be considered successful if the developed model has the ability to effectively balance reducing customer churn with maximizing revenue. The work that will be done developing and presenting this project is industry-relevant, and will prove highly valuable upon completion. One of the reasons for this is that customer retention is more cost-effective than acquisition. An empirical investigation conducted in 2016 published in the Journal of Marketing Research found that “a firm’s acquisition cost per customer is more sensitive to market position and competition than retention cost per customer” (Min et al., 2016). In addition to cost reduction, managing customer churn can also have an impact on a company’s revenue and profitability. As part of a 2024 literature review and bibliometric study conducted by Uday Bhale, a Doctor of Marketing from Lovely Professional University in India, and Harpreet Bedi, Professor at the Mittal School of Business, it was concluded that “a five percent churn reduction can improve a company’s profitability by up to twenty-five percent” (Bhale & Bedi, 2024). Fred Reichheld came to a similar conclusion in his article for Bain & Company titled, “Prescription for cutting costs” in 2001. A New York Times best-selling author, graduate of Harvard Business School, business strategist, and creator of the Net Promoter System of management, Fred poses, “Across a wide range of businesses, customers generate increasing profits each year they stay with a company. Return customers tend to buy more from a company over time. As they do, your operating costs to serve them decline. And they’ll often pay a premium to continue to do business with you rather than switch to a competitor with whom they’re neither familiar nor comfortable” (Reichheld, 2001). However, if a company ignores churn, their customers may choose to switch to competitors offering better pricing or services. That is why a project like this, which seeks to predict and reduce this churn can be so beneficial.

**Data Description**

The data utilized for this project can be found on Kaggle, but originates from IBM ([Data can be retrieved here](https://accelerator.ca.analytics.ibm.com/bi/?perspective=authoring&pathRef=.public_folders%2FIBM%2BAccelerator%2BCatalog%2FContent%2FDAT00148&id=i9710CF25EF75468D95FFFC7D57D45204&objRef=i9710CF25EF75468D95FFFC7D57D45204&action=run&format=HTML&cmPropStr=%7B%22id%22%3A%22i9710CF25EF75468D95FFFC7D57D45204%22%2C%22type%22%3A%22reportView%22%2C%22defaultName%22%3A%22DAT00148%22%2C%22permissions%22%3A%5B%22execute%22%2C%22read%22%2C%22traverse%22%5D%7D)), who has created a dataset for the company, Telco. The information collected in Excel is specifically for 7,043 California-based Telco customers, and occurs in Q3 (July-September). The dataset is comprised of several categories of information pertaining to the customers, including their demographics, their locations, the services that they are provided, and the status of the customers (their level of satisfaction, whether or not they have churned, etc.).

**Key Variables in the Data**

Gender – Categorical variable that tracks whether the customer is male or female.

Partner – Categorical variable that tracks whether the customer has a partner or not.

Dependents – Categorical variable that tracks whether the customer has dependents or not.

Number of Dependents – Discrete, numerical variable that tracks how many dependents a customer has.

Tenure Months – Discrete, numerical variable that tracks how many months a customer has been with Telco.

Phone Service – Categorical variable that tracks whether the customer has a phone plan through Telco.

Multiple Lines - Categorical variable that tracks whether the customer has multiple phone lines or not.

Internet Service - Categorical variable that tracks whether the customer has Internet service through Telco.

Online Security - Categorical variable that tracks whether the customer is enrolled in Telco’s online security service.

Online Backup - Categorical variable that tracks whether the customer is enrolled in Telco’s online data storage/backup service.

Device Protection - Categorical variable that tracks whether the customer has a device protection plan on a Telco provided device.

Premium Tech Support - Categorical variable that tracks whether the customer is enrolled in Telco’s premium tech support service.

Streaming TV - Categorical variable that tracks whether the customer streams TV as part of their Telco internet service.

Streaming Movies - Categorical variable that tracks whether the customer streams movies as part of their Telco internet service.

Streaming Music - Categorical variable that tracks whether the customer streams music as part of their Telco internet service.

Contract - Categorical variable that tracks the type of contract a customer is on. Options include month-to-month, one-year, or two-year.

Paperless Billing – Categorical variable that tracks if a customer is enrolled in paperless billing.

Payment Method - Categorical variable that tracks the payment method(s) that customers use. Included are both manual and automatic payment methods.

Monthly Charges – Continuous, numerical variable that tracks how much customers are charged per month.

Age – Discrete, numerical variable that tracks customer ages.

Referred a Friend – Categorical variable that tracks whether a customer has referred another customer to Telco.

Number of Referrals – Discrete, numerical variable that tracks how many customers a Telco customer has referred to the company.

Avg Monthly Long-Distance Charges – Continuous, numerical variable that tracks the average monthly long-distance charges a customer receives.

Total Long-Distance Charges - Continuous, numerical variable that tracks the total amount of long-distance charges a customer has received.

Avg Monthly GB Download – Continuous, numerical variable that tracks the average monthly gigabytes of data a customer has downloaded.

Unlimited Data – Categorical variable that tracks whether a customer is enrolled in an unlimited data plan.

Total Refunds – Continuous, numerical variable that tracks the total amount that a customer has been refunded.

Total Extra Data Charges - Continuous, numerical variable that tracks the total amount from data overages that a customer has been charged.

Total Revenue – Continuous, numerical variable that tracks the total amount of revenue that Telco has made on a customer.

Churn Label – The target variable of this dataset, that determines whether or not a customer has churned.

**Exploratory data analysis (EDA)**

EDA was started by first determining if all the entries in the dataset were unique, dropping any columns that would not be explored or used in the model, checking for null/missing values, and ensuring each variable was the correct data type. For numerical missing values, median was filled in rather than dropping the rows to retain as much information as possible. After working through these checks, EDA was split between exploring the numerical variables and the categorical variables. The summary statistics were checked for each numerical variable, in addition to distribution plots such as box-and-whisker plots and histograms. For categorical variables, pie charts were used to show value distribution for variables with more than binary values, such as Contract and Payment Method. For the others, “yes’s” and “no’s” were totaled to answer questions such as, “How many customers have dependents?”

After the variables were explored individually, the categorical and numerical variables were measured in how they correlated with churn. For numerical variables, this process involved finding the average values of each variable that was associated with churn (1) and not churning (0), as well as calculating each variable’s correlation with churn. After determining each variable’s correlation, the low and moderate correlating variables were transformed using their natural log, a MinMax scaler, and a standard scaler to see if any predictive power could be added to them, or to make them more useful. Through this process, it was found that the natural logs of Total Revenue, Total Charges, and Total Monthly Charges had higher correlations with churn than their original forms. For categorical variables, a series of tests were used, including a chi-square test to check for significant relationships between the variables and churn, and a Cramer’s V test to check the strength of the relationships found in the chi-square test. The final stage of EDA before moving forward into modeling was to create interaction terms between the categorical variables, and between the categorical and numerical variables. This was done in order to obtain any additional predictive power or information from the existing values. A method was created that took in the categorical and numerical columns and combined them together, giving them each names that depicted the relationship that each variable tried to capture, such as “Interaction\_Paperless Billing Unlimited Data\_No\_Yes”, which stands for a customer that is enrolled in an unlimited data plan that is not enrolled in paperless billing. In addition to providing potential predictive power, these interaction terms were created to assist in customer segmentation. If an interaction term was ruled significantly important in the results of a model, its interaction could provide insight into the types of customers more or less likely to churn.

**Modeling**

In the modeling stage, I built and tested models using four different machine learning classification algorithms in order to determine the best fit. The first algorithm, logistic regression, predicts the probability that an input will belong to either the positive class (1) or the negative class (0), rather than predicting the outcome directly. In the case of this problem, logistic regression will determine the likelihood that a customer will churn (1) or not churn (0), based on the information it is given. The second algorithm, scaled vector classifier, looks to separate data by the best boundary that divides and captures the two classes. The third algorithm, random forest, is an ensemble method, which means it aggregates multiple models together to produce a better result. In this instance, random forest combines several decision trees to make predictions. The fourth algorithm, extreme gradient boosting, or XGBoost, is an ensemble method similar to random forest, except that it builds decision trees sequentially rather than independently as random forest does. Sequentially built decision trees focus on correcting the errors made by previous trees to achieve a better result, whereas independently built decision trees are built at the same time, each with a different subset of the data, and their predictions are combined.

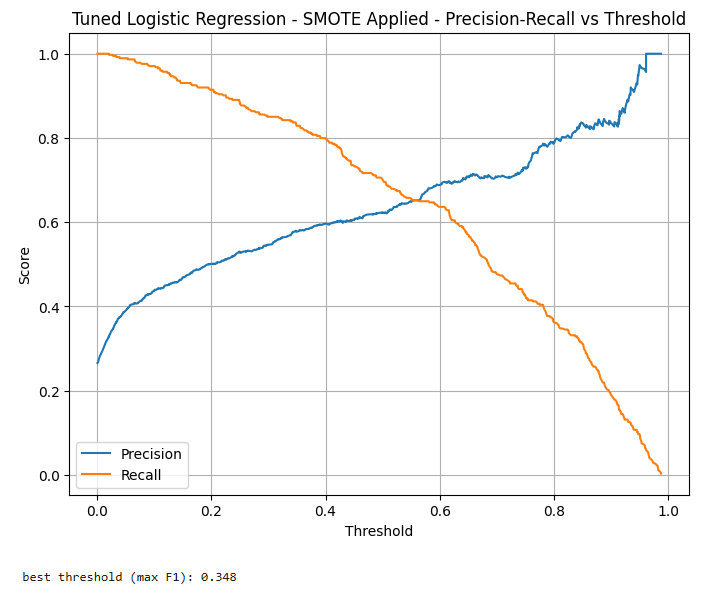
Each of these modeling approaches come with strengths and weaknesses. Logistic regression is the easiest of the four to interpret results from, outputting its probabilities directly to explain variable effects on prediction. It is also fast to run and predict on both small and large datasets. Logistic regression can be sensitive to multicollinearity, where two or more variables are highly correlated, and can make it difficult to determine the individual predictive power of your variables. It also struggles with highly imbalanced classes, such as the data being used in this project, so careful hyperparameter and threshold tuning, and resampling is usually required. Scaled vector classifier can be very effective with highly dimensional data (lots of feature variables), which can be beneficial for this dataset. However, the algorithm is very slow to train and predict on large datasets such as this and is arguably the hardest to interpret results from out of the four models. Random forest is robust against overfitting, which is when a model gets very good at memorizing training data, and does not generalize well on new, unseen data. It also is easy to determine feature variable importance from and can work well with nonlinear relationships. For large datasets like this though, random forest can perform slower than algorithms like logistic regression, and can be difficult to interpret on occasion, its decision-making process difficult to follow through the many trees. XGBoost is a very powerful algorithm that offers many benefits, including handling complex data patterns well and providing feature importances. However, XGBoost must be tuned carefully to avoid overfitting, and its results can be hard to interpret.

The models were evaluated on several performance metrics, including recall, F1-score, and AUC-ROC. Recall is the measure of the proportion of true positives and true negatives that a model correctly identifies. Class 1 recall was looked at in particular, as the data has a heavy imbalance towards class 0 (non-churners) over class 1 (churners). Recall is a very important metric to consider in this project, as false negatives (incorrectly identifying churners) are much more costly than false positives (incorrectly identifying non-churners). If the model wrongly predicts a customer that will not churn is likely to churn, the result would be taking action to try and retain that customer. This action could result in some lost revenue, however, the retention cost would be minimal compared to the cost of a lost customer. F1-score provides a balance of precision and recall, measuring the harmonic mean of the two metrics. Though recall is more important in this project than precision, giving up too many false positives to lessen the number of false negatives can hurt profits as well, so finding a good trade-off balance between the two metrics is key. AUC-ROC measures the overall performance of a classification model and depicts how well the model distinguished between the positive and negative classes. A model with an AUC score of 1 perfectly distinguishes between the two classes, whereas a model with an AUC score of 0.5 is no better than randomly guessing.

To begin modeling, the data was split into 80% training and 20% testing, and the values for the target variable, “churn”, were stratified, as the dataset has a heavy imbalance of non-churners (0) over churners (1). This was done to help ensure that the models tested would learn and generalize well on the test data. The variables in the dataset went through two different preprocessing transformations before they were tested in the models. For categorical variables, OneHotEncoding was used to convert the variables over to numerical seamlessly for utilization. For numerical variables, a StandardScaler was used. The StandardScaler reduces the impact of any extreme values present in the numerical columns, giving each column a mean of 0 and a standard deviation of 1 (a standard normal distribution). These two transformers were input into a column transformer to be put into a pipeline alongside the model(s) to be tested for repeatability, to reduce redundancy in the code, and to ensure the process was smooth and error free. The non-baseline models tested went through 5-fold, cross-validated grid searches, in order to determine the best values for each algorithm's hyperparameters. The best values were determined through average 'recall' scores, the main metric to compare model performance.

There was a total of twelve models that were built and tested during the modeling stage. The first four were baseline models of the four algorithms selected: logistic regression, random forest, extreme gradient boost, and scaled vector classifier. These were run using default parameters for each algorithm, and without any resampling techniques for class imbalance. The next four were the same models, but with their hyperparameters tuned through grid searches, and any available built-in class-weight balancing parameters. The final four models were hyperparameter tuned like the previous four, but with SMOTENC applied to the dataset. SMOTENC, or synthetic minority over-sampling technique for nominal and continuous, is a resampling technique that generates synthetic data for the minority class of a dataset, using “nearest neighbors.” This technique was used to oversample the number of churners (class 1) in the dataset to help deal with the class-imbalance. SMOTENC was used in particular over SMOTE, so that both the categorical and numerical variables could be used. The class distribution prior to SMOTENC was 5,172 non-churners (0) and 1,869 churners (1), and afterwords became 4,139 non-churners and 4,139 churners. To find the optimal threshold for each model, precision-recall vs threshold graphs were created to show how the values of precision and recall increase/decrease with a higher or lower threshold. These graphs were created with finding the threshold that resulted in the maximum f1-score in mind, with the goal of obtaining high recall scores with the highest precision available for each model.

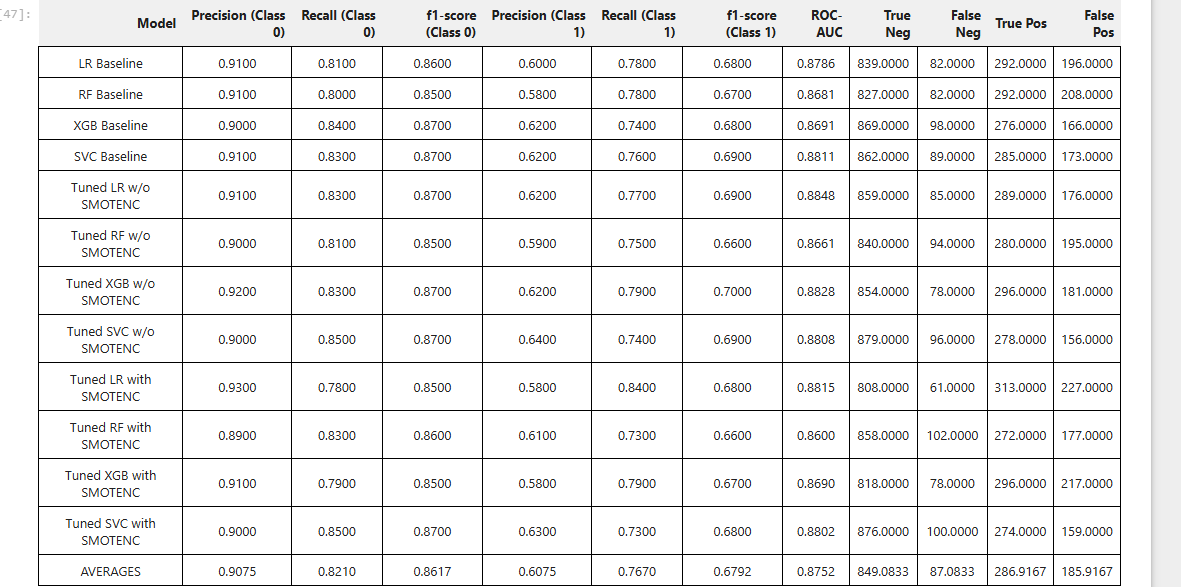
*Precision-Recall vs Threshold Graph for a Model – Tuned Logistic Regression with SMOTENC*



When selecting the best model, there were several criteria in mind that were considered for each choice. First of all, the optimal model needed to score a high level of recall, particularly class 1 recall, as incorrectly identifying churners as non-churners is more costly than incorrectly identifying non-churners as churners in this project. This should also come with a moderately high f1-score and AUC-ROC as well, proving the model’s ability to perform overall, and distinguish between the positive and negative classes. Second of all, the optimal model should be able to train and predict quickly, and not be too computationally expensive. Third of all, the results of the model should be easy to interpret and implement into Telco’s business strategy.

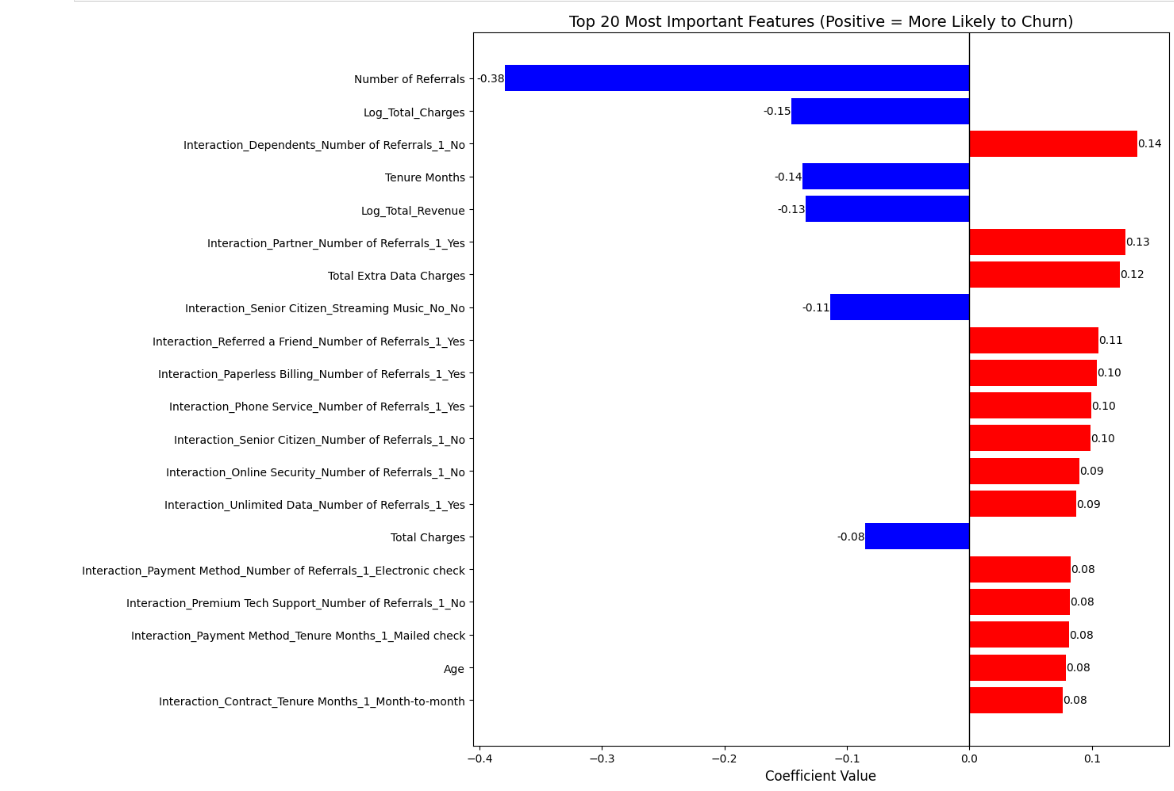
**Results**

*Performance Metrics for Each Model Tested, with Averages Included*

****

After considering the results of each model, it was decided that the model to move forward with is the Tuned Logistic Regression model with SMOTENC applied. From an interpretability standpoint, its results are the easiest to understand when compared to Random Forest, Extreme Gradient Boosting, and Scaled Vector Classifier, as the coefficients of the contributing feature variables have direct meaning. Each one represents the log-odds change in the odds of churn, provided the other variables hold constant. From a metrics standpoint, the model boasts the highest class 1 recall score of any of the models at 0.84, the third highest ROC-AUC score at 0.8815, and the fewest number of false negatives at only 61. This model also was tested with balanced data, albeit from SMOTENC. This likely means it is more accurate and less biased than the Tuned LR model with the original class distribution. The high number of false positives will be something to monitor, and the threshold may be changed in order to lessen the number of false positives at the result of increased false negatives when determining what makes the most sense profit wise. Logistic Regression also has a lower computational cost than XGB and SVC, and comparable cost to Random Forest.

*Feature Importances for the Chosen Model*



The above bar chart showcases the top twenty most important feature variables for the chosen model. When translated from coefficients, to log odds, to then actual odd percentage, the business interpretations for each of the features are as follows:

|  |  |  |
| --- | --- | --- |
| **Feature Variable** | **Coefficient to odds %** | **Translation** |
| Number of Referrals | -0.38 = -31.6% change | Each additional customer referral decreases the odds of churn by 31.6%. |
| Log\_Total\_Charges | -0.15 = -13.9% change | Higher total charges decrease the odds of churn by 13.9%. |
| Interaction\_Dependents\_Number of Referrals\_1\_No | 0.14 = 15.1% change | Customers with no dependents and 1 customer referral have an increased odds of churn by 15.1% |
| Tenure\_Months | -0.14 = -13% change | Each month of tenure for a customer reduces the odds of churn by 13%. |
| Log\_Total\_Revenue | -0.13 = -12.2% change | Higher total revenue from a customer decreases the odds of churn by 12.2%. |
| Interaction\_Partner\_Number of Referrals\_1\_Yes | 0.13 = 13.9% change | Customers with partners who have made 1 referral have an increased odds of churn by 13.9%. |
| Total Extra Data Charges | 0.12 = 12.7% change | Higher total extra data charges increase the odds of churn by 12.7%. |
| Interaction\_Senior\_Citizen\_Streaming Music\_No\_No | -0.11 = -10.5% change | Customers who are not senior citizens (Under 65) that do not stream music are 10.5% less likely to churn. |
| Interaction\_Referred a Friend\_Number of Referrals\_1\_Yes | 0.11 = 11.6% change | Customers who have referred a friend, and made only 1 referral are 11.6% more likely to churn. |
| Interaction\_Paperless Billing\_Number of Referrals\_1\_Yes | 0.10 = 10.5% change | Customers who are enrolled in paperless billing, and have referred 1 customer are 10.5% more likely to churn. |
| Interaction\_Phone Service\_Number of Referrals\_1\_Yes | 0.10 = 10.5% change | Customers that have phone plans and have made 1 referral are 10.5% more likely to churn. |
| Interaction\_Senior Citrizen\_Number of Referrals\_1\_No | 0.10 = 10.5% change | Customers who are not senior citizens that have made 1 referral are 10.5% more likely to churn. |
| Interaction\_Online Security\_Number of Referrals\_1\_No | 0.09 = 9.4% change | Customers who do not have online security who have made 1 referral are 9.4% more likely to churn. |
| Interaction\_Unlimited Data\_Number of Referrals\_1\_Yes | 0.09 = 9.4% change | Customers who have unlimited data plans who have made 1 referral are 9.4% more likely to churn. |
| Total Charges | -0.08 = -7.7% change | Similar conclusion to Log\_Total\_Charges. |
| Interaction\_Payment Method\_Number of Referrals\_1\_Electronic check | 0.08 = 8.3% change | Customers who pay with electronic check and have referred 1 customer are 8.3% more likely to churn. |
| Interaction\_Premium Tech Support\_Number of Referrals\_1\_No | 0.08 = 8.3% change | Customers who do not have premium tech support who have referred 1 customer are 8.3% more likely to churn. |
| Interaction\_Payment\_Method\_Tenure Months\_1\_Mailed check | 0.08 = 8.3% change | Customers that have been with Telco for 1 month, that pay through mailed check are 8.3% more likely to churn. |
| Age | 0.08 = 8.3% change | Older customers have an increased odds of churn by 8.3%. |
| Interaction\_Contract\_Tenure Months\_1\_Month-to-month | 0.08 = 8.3% change | Customers who have been with Telco for 1 month, that are on Month-to-month contracts, are 8.3% more likely to churn. |

**Conclusion**

Based on the results of the model, there are three different customer profiles that Telco should be aware of in planning their business strategy to reduce churn. These three are:

High Risk Customers

* Customers that are on month-to-month contracts and/or new customers.
* Manual paying customers (mailed check & electronic check).
* Single referral customers that do not have deep product/service adoption.
* Customers without premium support or online security services.
* Customers that cause extra data charges.

For these customers, my recommendations are to offer 3–6-month discounted bundles upon customer sign up in order to lock in these high-risk customers early. Autopay services should be incentivized, potentially offering customers discounts on their bill for enrolling in autopay versus paying manually. Offer new customers 30-day free trials of their premium tech support and online security services upon sign up. Referrals should also be incentivized, offering referral bonuses to customers, or offering months of free services like premium tech support and online security upon achieving 2 or more referrals.

Moderate Risk Customers

* Single service customers (only on phone plans, or TV plans).
* Customers that are enrolled in paperless billing and have low engagement.
* Senior citizen customers without support packages (premium tech support, device protection, online security, etc.)

For these customers, my recommendations are to create bundles at a discount, giving them the opportunity to adopt services such as internet, TV, and premium tech support at a lower cost, making them further engaged and more likely to stay on board. Discounts could be created on services like device protection and online security for customers that stick with paperless billing, creating a form of a “digital loyalty” program. For senior citizen customers specifically, a “Senior Care” program could be created, that offers these customers easy access to tech support hotlines and chatbots, customer service, and discounts on services like premium tech support and online security to assist them with service adoption.

Low Risk Customers

* Customers that are high total spenders.
* Customers that have multiple referrals (2+).
* Customers with longer tenure.

These customers in particular should be recognized for their loyalty to Telco. VIP programs should be created that give these long tenured, multiple-referral customers exclusive offers, loyalty credits on bills, early access on deals for new devices, and “thank-you” gifts for customers that reach specific tenure milestones, such as 1 year, 5 years, and 10 years.

Moving forward, Telco should monitor these customer segments over their next quarter, noting the effectiveness of the retention recommendations. If higher risk customers have become moderate or lower risk, then the strategies are working, and should be reduced in intensity so that unnecessary levels of retention costs aren’t achieved. If moderate risk customers have become high risk, respond immediately with stronger offers so that these customers are not lost. If low risk customers have moved to moderate or even high risk, reach out to these customers through check-in emails and try to deduce where the dissatisfaction is coming from so that it can be appealed.

After each quarter, the model should be refreshed and validated. Changes in customer behavior, pricing changes, or other outside factors may have affected model performance, and therefore, should be reevaluated using new data. The model could potentially be under or over-estimating churn likelihoods for customers. If performance is down, retrain the model using the most recent data and churn outcomes, and determine if new feature variables become important/less important. The optimal threshold may have changed as well, and needs retuned based on the new information. This continuous improvement cycle will ensure that Telco has the most up-to-date, accurate model to draw insights and craft business strategy from.

**References**

Bhale, U.A., Bedi, H.S. (2024). *Customer Churn Construct: Literature Review and Bibliometric Study*. Management Dynamics: Vol. 24: No. 1, Article 1. [https://managementdynamics.researchcommons.org/cgi/viewcontent.cgi?article=1327&c ontext=journal](https://managementdynamics.researchcommons.org/cgi/viewcontent.cgi?article=1327&c%09ontext=journal)

Min, S., Zhang, X., Kim, N., & Srivastava, R.K. (2016). *Customer Acquisition and Retention Spending: An Analytical Model and Empirical Investigation in Wireless Telecommunications Markets*. Journal of Marketing Research, 53(5), 728-744. <https://doi.org/10.1509/jmr.14.0170>

Reichheld, F. (2001). *Prescription for cutting costs*. Bain & Company. <https://media.bain.com/Images/BB_Prescription_cutting_costs.pdf>