Churn Analysis of Software-as-a-Service Company

Introduction

“Customer service shouldn’t just be a department; it should be the entire company.” As Tony Hsieh, former CEO of Zappos, had already stated in the early 2000s, customer satisfaction is crucial for any business that operates in a competitive environment. Dissatisfied clients churn from company’s customer base, which directly impacts revenue and long-term growth prospects. Preserving and expanding existing customer relationships is more cost-efficient than new customer acquisition. Research has shown that acquiring a new customer is 5 to 25 times more expensive than retaining an existing one (Gallo, 2014) and reducing churn by just 5% can increase profits by more than 25% (Reichheld and Schefter, 2019). High churn rates can stifle a company’s growth and make it more difficult to expand the customer base and to cross- and upsell products to the existing one. By understanding why customers leave, businesses can innovate and improve their offerings, address the needs and preferences of their target group and adapt to changing market conditions.

In this report, we explore potential contributors to client churn in a subscription-based business, apply some machine learning approaches in forecasting, and end with actionable insights and recommendations for the company.

Data Overview

The data set referenced for the remainder of this report includes 505,206 records of customer churn outcomes, each with 11 associated attributes. The attributes include customer ID, age, gender, tenure, usage frequency, support calls, payment delay, subscription type, contract length, total spend, and last interaction. Customer ID represents a numerical value, but upon consolidating the provided testing and training CSV files, it was evident that this might not be consistently applied or maintained, so it was excluded. Age is given in years (integer), gender is male or female (binary string), tenure is given in months (integer), usage frequency is the number of days the client used the product in the past 30 days (integer), support calls are the number of calls made to the company’s support centre in the past 30 days (integer), payment delay is the number of days the client is past due (integer), subscription type includes basic, standard, and premium (strings), contract length includes annual, quarterly, and monthly (strings), total spend is the client’s amortized spend in the most recent month, regardless of contract length or payment terms (float), and last interaction represents days since logging into the product (integer). Churn is binary (1 = churn, 0 = retained).

An important additional metric will be referenced throughout this report: retention rate. This is calculated as 1 – sum of churn / count of included records, and it will be sliced in a variety of ways. For example, to assess the retention rate for males over 50 years old, we calculate this as 1 - (sum of churn where gender = male and age > 50) / (count of records with churn outcomes where gender = male and age > 50). The churn rate is simply the inverse of this (where retention rate is 80%, churn rate would be 20%).

The data was obtained from Kaggle (Shahid Azeem, n.d.) and was assigned a 10.00 usability rating. The data is described as an anonymised customer churn data set, but no firmographic information is provided (industry, firm size, etc.). Additionally, there is no time-period specified, so the data is assumed to provide aggregate information for some fixed period.

Exploratory Statistical Analysis

The following are the summary statistics for the 505,206 observations in the dataset:

statistic	Age	Tenure	Usage Freq.	Supp. Calls	Payment Delay	Total Spend	Last Interaction
Mean	39.7	31.4	15.7	3.8	13.5	620.1	14.6
Median	40	32	16	3	13	648.9	14
Standard deviation	12.7	17.2	8.6	3.1	8.5	245.3	8.6
Minimum	18	1	1	0	0	100	1
Maximum	65	60	30	10	30	1000	30

Table 1

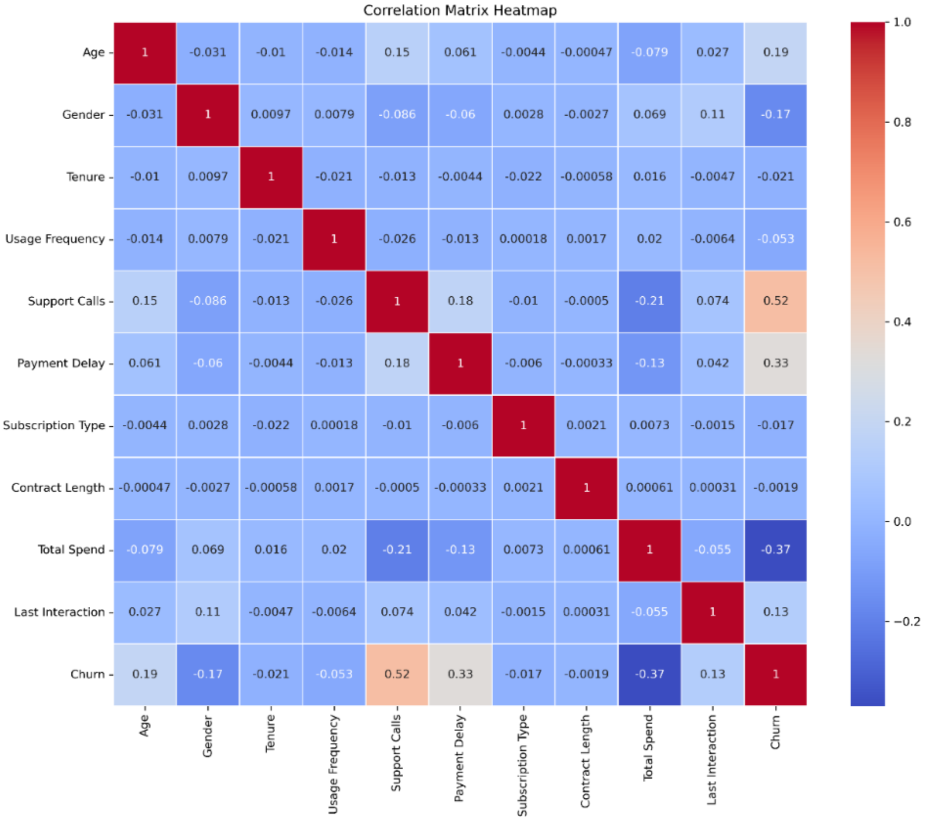
The distributions of the attributes in the dataset generally have normal distributions, with the exception of support calls. The higher average mean of support calls than the median indicates that a few customers made a lot more calls than others. A few interesting observations on the distribution of the data are summarized below and the full distributions are visualised in the accompanying dashboard:

- **Age**: The average customer age is approximately 40 years old, with a normal distribution from 18 to 65. Retention increases from 45% for clients aged 18 to 25 to 60% for ages 40-50, past which it declines sharply to below 10%. The 51+ age group had more monthly contracts and support call volumes.
- **Gender**: when looking at retained customers specifically, only 35% are female. There are 55,340 more male customers than female customers in the dataset.
- **Support Calls**: retained clients made far fewer support calls than their churned counterparts. Clients with fewer than 4 support calls retained at 60-70%, with 4 at 45%, with 5 at 13%, and beyond that at 8-9%.

- **Total Spend**: Retained clients aged 51 or older had a 23% lower price than the rest of the client base. Retained clients spent 34% more than churned clients. Clients below £500 in spend had retention rates around 10%, while above that threshold, the rate jumps to nearly 60%.
- **Payment Delays**: beyond 20 days, retention rate declines from 50%+ to under 6%.
- **Tenure** and **Subscription Type** have less material movement in retention rates.

Correlation studies and logistic regression – Measuring the influence of different factors on churn rate

To understand the relationships between multiple variables, we calculated correlation pairwise and plotted the result. As seen in **Figure 1**, there is a moderately strong positive correlation between support calls and payment delay with churn, indicating that customers that call often or fail to pay their fees are more likely to churn. On the other hand, customers who spend more in total, are less likely to churn. This is plausible: A customer spending more tends to make use of the services provided and therefore stays loyal. Age seems to also positively correlate with support calls (meaning that older customers might need more support) and churning.



	Age	Gender	Tenure	Usage Frequency	Support Calls	Payment Delay	Subscription Type	Contract Length	Total Spend	Last Interaction	Churn
Age	1	-0.031	-0.01	-0.014	0.15	0.061	-0.0044	-0.00047	-0.079	0.027	0.19
Gender	-0.031	1	0.0097	0.0079	-0.086	-0.06	0.0028	-0.0027	0.069	0.11	-0.17
Tenure	-0.01	0.0097	1	-0.021	-0.013	-0.0044	-0.022	-0.00058	0.016	-0.0047	-0.021
Usage Frequency	-0.014	0.0079	-0.021	1	-0.026	-0.013	0.00018	0.0017	0.02	-0.0064	-0.053
Support Calls	0.15	-0.086	-0.013	-0.026	1	0.18	-0.01	-0.0005	-0.21	0.074	0.52
Payment Delay	0.061	-0.06	-0.0044	-0.013	0.18	1	-0.006	-0.00033	-0.13	0.042	0.33
Subscription Type	-0.0044	0.0028	-0.022	0.00018	-0.01	-0.006	1	0.0021	0.0073	-0.0015	-0.017
Contract Length	-0.00047	-0.0027	-0.00058	0.0017	-0.0005	-0.00033	0.0021	1	0.00061	0.00031	-0.0019
Total Spend	-0.079	0.069	0.016	0.02	-0.21	-0.13	0.0073	0.00061	1	-0.055	-0.37
Last Interaction	0.027	0.11	-0.0047	-0.0064	0.074	0.042	-0.0015	0.00031	-0.055	1	0.13
Churn	0.19	-0.17	-0.021	-0.053	0.52	0.33	-0.017	-0.0019	-0.37	0.13	1

Figure 1

Logistic regression is a useful statistical method in this context to model the probability of customers churning: The value of each independent variable helps us to quantify how each independent variable affects the probability of the outcome allowing to determine which factors are the most influential for predicting events such as customer churn. We can also use it to predict the likelihood of an individual customer churning and identify high-risk customer to target with retention strategies.

A logistic regression model was calculated using the sklearn kit in Python. The result is summarized in **Table 2**.

With a Pseudo R^2 value of 0.3809, 38.09% of the variance in churn can be explained by the coefficients. Except for contract length, all coefficients are highly significant. The sign of the coefficient indicates if it increases or decreases the probability of churning. Tenure, usage frequency, subscription type and total amount spend decrease churning rate, meaning an older contract, the more it is used and the more exclusive the subscription type is, the less likely it is to be cancelled. On the other hand, if a customer is older, has more payment delays or needs to frequently call for support, the higher the risk of contract termination is. Males have a lower chance of churning as seen by a negative coefficient.

Factor	Coefficient	Standard Error	z-value	P>|z|	95% CI (lower bound)	95% CI (upper bound)
const	-0.957	0.023	-41.666	0.000	-1.002	-0.912
Age	0.025	0.000	78.380	0.000	0.024	0.025
Gender	-0.762	0.008	-97.136	0.000	-0.778	-0.747
Tenure	-0.002	0.000	-6.976	0.000	-0.002	-0.001
Usage Frequency	-0.013	0.000	-29.153	0.000	-0.014	-0.012
Support Calls	0.422	0.002	275.300	0.000	0.419	0.425
Payment Delay	0.086	0.000	174.148	0.000	0.085	0.087
Subscription Type	-0.038	0.005	-8.103	0.000	-0.047	-0.029
Contract Length	-0.005	0.004	-1.268	0.205	-0.014	0.003
Total Spend	-0.003	0.000	-190.911	0.000	-0.003	-0.003
Last Interaction	0.034	0.000	75.408	0.000	0.033	0.035

Table 2: Coefficients and statistical testing of the logistic regression model

For a more intuitive interpretation of the coefficients, we can calculate the odds ratios using this formula: $odds\ ratio = e^{coefficient}$

An odds ratio of > 1 means that the predictor is associated with a higher probability of the outcome occurring, whereas a ratio of < 1 indicates lower probabilities.

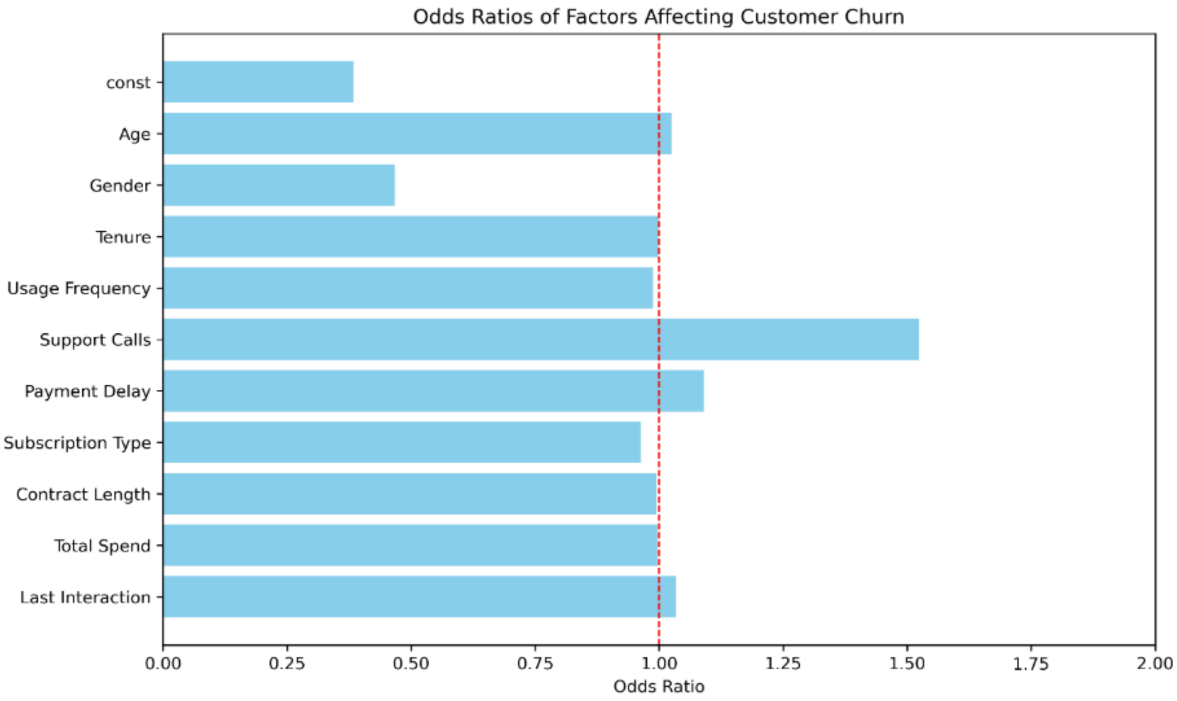


Figure 2

As seen in **Figure 2**, a customer calling for support is almost 50% more likely to churn, and males retain at higher rates. This has two major implications: the company needs to explore the reasons behind these trends and consider tailored offerings and enhanced customer support or user interface improvement.

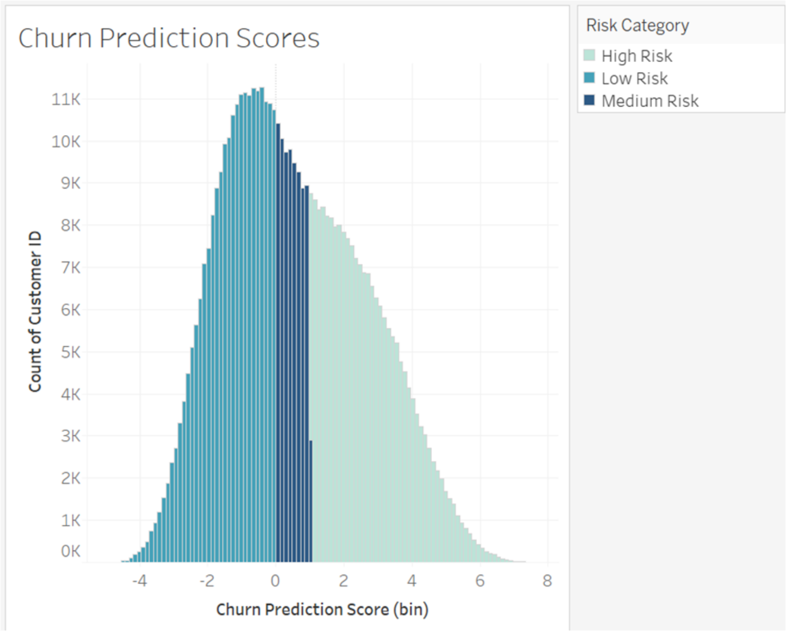


Figure 3

Leveraging the above regression model, we created a histogram of predicted churn scores (**Figure 3**). It has a mostly normal distribution. The calculated prediction score combines the effects of the independent variables, with frequencies displayed. The peak of the histogram lies between 0 and -2, indicating that the largest share of clients fall there. Customers with positive scores are at medium or high churn risk.

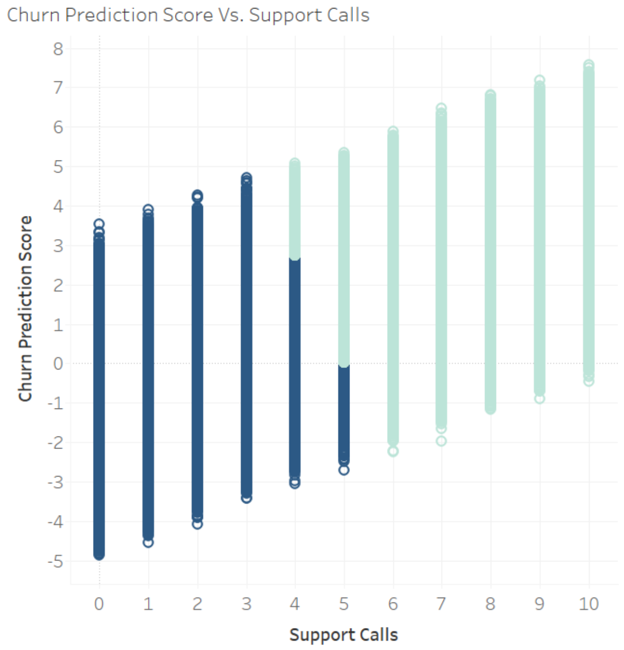


Figure 4

Negative scores (-4 to 0) imply a lower probability of churn. We would expect the low-risk group to include younger clients with fewer support calls, higher spend, more timely payments, and longer tenure. The implications of this are discussed in more detail in the “Action Plan” section. **Figure 4** illustrates the relationship between support call volume and churn. As the number of support calls increases, the likelihood that the customer churns also increases. When applying the cluster functionality in Tableau, two customer segments were identified, which can inform different recommendations.

Machine Learning Model – Decision Tree

Leveraging the statistical findings discussed above, the next step toward extrapolating findings to predict churn is implementing machine learning techniques. Some of the options considered included k-means clustering, Naive Bayes, logistic regression, decision trees, random forests, and gradient boosting. While k-means could help the company to categorize their client base, it would need to be combined with other models to generate an ultimate churn prediction. Naive Bayes assumes feature independence, which cannot be assumed here. For example, total spend and support calls have a correlation of –0.21. Last interaction and gender have a correlation of 0.11. These are weak correlations, but we have already identified some limitations in the data set. As the company ingests more data points into this model, it is possible that new, stronger correlations surface. The more attributes added pertaining to this company, the less likely feature independence becomes. Logistic regression was performed and discussed above. This method is limited by the existence of nonlinear relationships among attributes, particularly due to the presence of binary and other categorical variables. It provides useful insights around to what level additional attributes may be needed, however. Decision trees offer clear visualizations around nodes and paths, as well as interpretable feature importances. Non-linear and mixed data types are less of an issue in this methodology, and performance is quite fast. It could result in overfitting, which could be improved by a random forest model (multiple decision trees, with majorities applied to the final outcome) or gradient boosting (also multiple decision trees, but in sequence to improve iteratively).

Due to the state of this company's data, the selected model was a single decision tree. This will allow the company to identify immediate, tangible conclusions and leverage this to enhance the data pipeline itself. Future iterations should consider random forests and gradient boosting, when more time series data and additional attributes are considered, in more meaningful ways. This will be explored further in the "Action Plan" section, below.

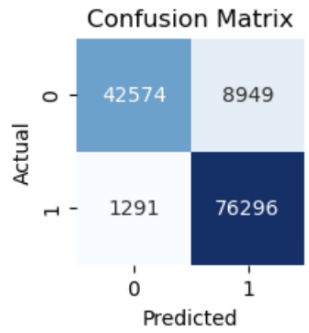
Simply plugging all possible attributes into a standard decision tree model did not produce very high accuracy. Therefore, we implemented several transformations of the attributes to align with findings from the statistical exploration. Significant patterns were observed in clients over 50 years of age, so this was converted to a binary variable. The same was appropriate for payment delays (<=20 vs. >20), total spend (<500 vs. >=500), last interaction (<=15 vs. >15), and contract length (monthly versus non-monthly, such that annual and quarterly were grouped together). Usage frequency and support calls were each placed into three buckets (<=2, 3-9, 10+ for the former, and <=2, 3-4, 5+ for the latter). Tenure and subscription type were not deemed important, which suggests potential opportunities in the company's pricing model, discussed in "Action Plan" below.

The model output is shown in **Table 3** below.

	Precision	Recall	F1 Score	Support
Retained	0.97	0.83	0.89	51,523
Churned	0.90	0.98	0.94	77,587
Accuracy	n/a	n/a	0.92	129,110
Macro Avg	0.93	0.90	0.91	129,110
Weighted Avg	0.93	0.92	0.92	129,110

Table 3

With an accuracy of 92%, this model appears to perform quite well. However, a higher recall for churned clients versus retained clients indicates that we are worse at predicting retention. This is also visualised in the confusion matrix below.



The top row represents actual retention, with support of 51.5k. 8.9k was incorrectly predicted to churn, or 17.4%, which aligns with the 83% recall rate in the table above. Of the second row, only 1.3k of 77.6k churned clients (1.7%) were predicted to be retained (98% recall). Of the 85.2k predicted to churn, 76.3k (89.5%) did, and of the 43.9k predicted to be retained, 42.3k (97.1%) were. These align with the precision metrics in the model output. From the standpoint of a business that wants to operate conservatively, it is generally better to have positive "surprises". With this model, the business might over-predict churn, resulting in higher than forecasted revenue, ceteris paribus. This could help the company to avoid layoffs in unfavourable economic times; if there is higher churn than predicted, shortfalls in collections could prevent the company from meeting payroll obligations.

Feature importance is another insightful output of decision trees and is shown in **Table 4**.

Feature	Support Calls	Total Spend	Contract Length	Payment Delay	Age	Last Interaction	Gender	Usage Frequency
Score	0.411	0.230	0.139	0.101	0.073	0.040	0.005	0.002

Table 4: Gini Impurity – Feature Importance Scores

"Support calls" are the most important feature, with a score of 0.41. This means that this attribute represents 41% of the reduction in impurity caused by all attributes together. It is the most important decision-making factor in this tree. Total spend follows with a score of 0.23, with contract length at 0.14, payment delay at 0.10, age at 0.073, gender at 0.0047, and usage frequency at 0.0015. Removing gender and usage frequency did not improve model accuracy, so they have been left in.

We ran the same model with a random forest classifier for the decision tree classifier. The recall, precision, and accuracy stayed consistent; only the feature importance distributed some of the support call score to other features (following the above order, scores became 0.35, 0.19, 0.13, 0.14, 0.11, 0.043, 0.039, 0.0025). Most notably, payment delay jumped ahead of contract length, but not by a large amount. The story is unchanged, and there is more work to do to enhance the robustness of the data pipeline before fine-tuning model methodologies. Hyperparameter tuning is also less relevant with the current data available, as more features would be useful here.

Action Plan: Insights and Recommendations

Data Integrity Project

In the analysis of this data, several potential data improvements came to light. Customer IDs were removed due to their apparent inconsistency. It could be that there was an issue in the original splitting of the training and testing data sets that did not correctly account for unique Customer IDs. When merged, there are some duplicates with different attribute values. It is unclear if a paying customer can have shared accounts (where customer ID would be the same, but sub-accounts can churn from the main account), or if it is a data integrity issue. Basic, Standard, and Premium subscriptions are not priced very differently, discussed above, but is the company confident that these figures are correct? The training data set had all clients over 50 years of age churning, while the testing set had only around half churning. For this analysis, we assumed the training and testing data was not split appropriately, so we merged these and re-split them randomly for the decision tree. However, it is possible that there is some data anomaly in the 51+ age group. Additionally, some clients with an “Annual” contract length are churning before a tenure of 12 months. The contract terms may be too lenient and allow termination for convenience, discussed in the pricing model section below. However, we must also consider the possibility of data inaccuracies here.

In addition to investigating data inaccuracies and anomalies, more data is needed for a more robust data pipeline. An important conclusion of the logistic regression section is that our current attributes only explain less than 40% of the variance in churn. An expanded data set will not only enhance churn predictions generated by the machine learning model, but it will also allow the company to make more real-time, data-driven decisions. Time series data, even just aggregated at the monthly level, would allow the company to explore changes over time. The client lifecycle is a critical area of analysis for any subscription-based business. Did the client downgrade from Premium to Standard before churning, suggesting the Premium suite might not be as premium as the company intended? What are the trends in usage frequency, and could this uncover churn risk or upsell opportunity? When was the last interaction with the product, and can the customer service team reach out proactively? What have been the trends in total spend? Has the client negotiated payment terms down from annual or quarterly to monthly, and is payment delay increasing, suggesting possible credit risk in the client? Are support calls increasing despite consistent usage, and is this happening in aggregate across the client base, indicating an overarching product problem?

This last question highlights another key benefit of time series benefit by client: not only looking at client trends individually, but on a cohort basis (available already via “Tenure”) or other form of aggregation. Trends over time could be assessed by industry of the client, their Gartner type (Enterprise vs. Mid-Market), or other firmographic factors. Macroeconomic conditions could be explored as well; for example, were there different trends with hospitality clients during the pandemic?

This project would require involvement from the teams responsible for business systems and the data warehouse, provided the company has these. Subject matter experts (SMEs) would need to inform the definitions and application of any data added. They should be consulted as the machine learning predictions are implemented, to identify any churn rationale not captured by the model. It’s possible that there will always be factors impossible to predict in advance, but with cross-functional collaboration, the integrity and relevance of the data could be greatly improved.

Support Centre and Product Performance Assessment

In the decision tree model implemented above, support calls stood out as the most important feature. Fewer support calls were associated with increased retention; however, without business context, it is unclear whether the support centre is ineffective, or serious product issues beyond the scope of the support centre need to be investigated. It is also unclear whether each call went through, so the inability to reach a representative could be a driver of churn.

The company should view support calls over time to isolate whether the sheer volume of support calls is associated with churn as much as increases or decreases in those volumes over time. More data should be collected around the length of support calls and reasons for those calls. Any other avenues of customer support should be recorded as well. With more client information, more associations should be explored to see if the product is underserving certain client groups.

Pricing Model Overhaul

The lack of material differences in retention rate by tenure and subscription type, coupled with unclear pricing differentiation among the three subscription types, suggests opportunities to revisit the company's pricing model. The average spend by these types are not vastly different, so the product marketing team should revisit product packaging. Monthly contract lengths have 14% lower spend on average than annual and quarterly, regardless of subscription type. Typical subscription models include discounts for annual payments, as the company can fund its business via deferred revenue. Retained clients spent 34% more than churned clients, so there does appear to be some flexibility to revisit the pricing model and find ways to drive more value.

The company can also consider implementing payment penalties or usage incentives. The retention rate for clients drops sharply after crossing the 20-day payment delay mark. If the client has not paid, and they have not used the product, it is highly likely they will churn. Buttoning up contract terms could hold clients more accountable and drive engagement, reducing churn. There is a balance to strike between this and leaving enough leniency to bring clients onboard efficiently. Any product issues would need to be resolved before going in this direction.

Conclusion

In conclusion, our analysis of customer churn yielded several key takeaways. Retention is impacted by multiple variables including support calls, total spend, contract length, payment delays, and age. Clients older than 50 had lower retention and different customer support needs than younger clients. In general, the more support calls were made by a client, the more likely they were to churn. This suggests that investing in a robust customer service infrastructure and/or enhancing the product roadmap could potentially lead to long-term customer retention.

The team recommended three action plans to address some of these findings. Firstly, the predictive analysis would benefit from an expanded data set, including temporal data to capture the lifecycle of a client. Next, the company should evaluate avenues to increase customer satisfaction, particularly for those with multiple support calls in the past month. Finally, it should establish consistency in pricing models and create value and differentiation in the levels of services offered. In implementing these suggestions, the company can improve retention and offset product and customer services investment via a well-executed pricing model, enhancing profitability.

References

Shahid Azeem, M. (n.d.). Customer Churn Dataset. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset.

Gallo, A. (2014). The Value of Keeping the Right Customers. [online] Harvard Business Review. Available at: https://hbr.org/2014/10/the-value-of-keeping-the-right-customers.

Reichheld, F. and Schefter, P. (2019). The Economics of E-Loyalty. [online] HBS Working Knowledge. Available at: https://hbswk.hbs.edu/archive/the-economics-of-e-loyalty.