Customer Churn Analysis Report

Chang Shu, Iuliia Oblasova, Yiran Chen

1. Motivation for the Project

The motivation for the project is to evaluate the effectiveness of a discount market offer to decrease the probability of customers opting out from a telecommunication service provider. A discount offer is a service of value companies offer customers to improve customers' engagement. It is common practice for companies to increase their customer base by acquiring new customers as well as retaining the old ones. Commonly referred to as customer churn, the telecommunication service companies in particular suffer from a loss of valuable customers. [1] The cost of churn includes both lost revenue as well as the marketing costs involved. Therefore, investigating how a discount offer contributes to suppressing customer churn would provide valuable insights on marketing strategy for tele companies to combat churn.

2. Motivation for the Research Design being Used

Given the context of the study, the treatment and control group are defined as customers received with a discount offer and those who did not. In order to separate out the causal effect of a discount market offer, both treatment and control groups need to be ensured with the same potential outcomes. As the potential outcomes are fundamentally unobservalue, we will examine carefully on how the two different groups were being assigned and whether that process may have resulted in the groups being different in a way that is likely to affect the outcome churn probability. To adequately control for these differences, we will use propensity score matching to assure that treatment and control groups are well balanced in the chosen covariates except the treatment variable. The covariates of interest include additional services of a phone plan such as streaming TV access, streaming music, unlimited internet access, etc. Before matching, an initial check of logistic regression was conducted to identify possible statistically significant confounders as part of exploratory data analysis. After matching, we will be able to perform regression and establish correlations which imply causation by analyzing the coefficient for the treatment variable with statistical significance indicator.

3. Study Details and Data Preprocessing

The first set of data is a fictitious dataset obtained from IBM Cognos Analytics sample datasets on Telco customer churn. [2] The dataset includes four types of information. The first is demographic information from a representative sample of the customer population that have a service contract with the telecompany. The covariates include age, gender, address locations, etc. The second is information about the types of services each customer has, including contract type, number of months using a service, etc. These indicators are a measure of how long the customer has stayed with the service provider, which might be highly correlated with their decision to churn or not. The dataset also contains customer paying information like monthly charges, types of services each customer has, etc., which may not be directly linked to our outcome but included for further analysis. Lastly there is information on customers' level of satisfaction which is directly related to whether customers decide to churn or not. The above information is merged via 7043 unique customer IDs and unnecessary or duplicating columns are dropped.

One aspect to be noted is that our available dataset does not contain income information of a customer. However, income can be one of the most important influencing factors for a customer to decide to take up a discount offer or not, which would lead to selection bias for treatment assignment. To overcome this disadvantage, a second dataset of average income aggregated at City level from US Census Bureau 2018 American Community Survey 5-Year-Estimates is incorporated. [3] The data was cleaned by comparing and fixing the city names with those from the previous dataset, and then inner joined with the previous dataset by city names, leaving 5942 records in the end. By integrating customer income information, we are able to control the income difference between treatment and control group to minimize its effect on treatment assignment as well as outcome churn probability.

The final merged dataset removed unnecessary and redundant columns. The often-occuring "Yes/No" binary responses across different variables were modified to 1 and 0. For multi-category variables, dummy variables are created for each subcategory with 0 or 1 input. As the numeric variables have different scales, normalization is performed on each numeric value by subtracting the minimum value of that predictor and then divided by the gap from maximum to minimum value of that predictor. All numeric variables are adjusted to around the same scale this way. Regarding missing values, there is around 73% of missing data out of total records in Churn Category and Churn Reason. As the main project objective is to see what causal factors would lead to higher or lower probability of customer churn, the two variables were only used during Exploratory Data Analysis and removed for further modeling.

4. Exploratory Data Analysis

a) Geographic locations of users

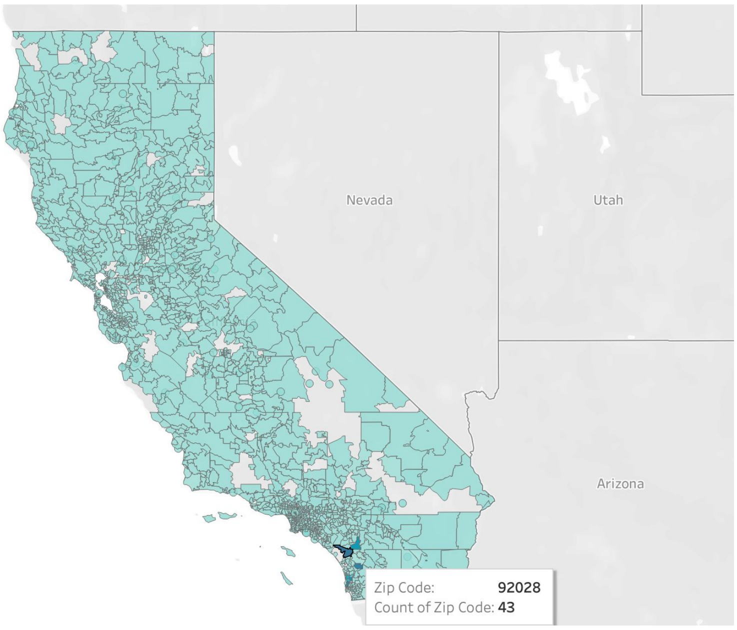
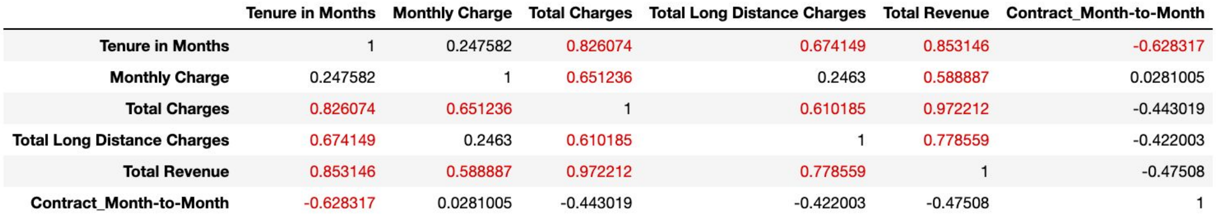


Figure 1. Geographic location of users

All of the users are located in California. Geographic unit is a Zip Code region with at least four respondents per unit on average. Whereas the majority of respondents are distributed uniformly across zip code areas, some of the units have a higher number of participants. The maximum number reported is 43 respondents in Zip Code 92028 located in Fallbrook, CA.

b) Multicollinearity

Because we are aiming to use linear regression, multicollinearity is an important factor to consider to ensure that features are independent. As displayed in the table below, some of the features are highly correlated. We will include only one feature from each pair to eliminate multicollinearity effect. Using 0.7 as the covariance cutoff value, Total Charges was removed in the later modeling process.



	Tenure in Months	Monthly Charge	Total Charges	Total Long Distance Charges	Total Revenue	Contract_Month-to-Month
Tenure in Months	1	0.247582	0.826074	0.674149	0.853146	-0.628317
Monthly Charge	0.247582	1	0.651236	0.2463	0.588887	0.0281005
Total Charges	0.826074	0.651236	1	0.610185	0.972212	-0.443019
Total Long Distance Charges	0.674149	0.2463	0.610185	1	0.778559	-0.422003
Total Revenue	0.853146	0.588887	0.972212	0.778559	1	-0.47508
Contract_Month-to-Month	-0.628317	0.0281005	-0.443019	-0.422003	-0.47508	1

Figure 2. Collinearity of Covariates

c) Reasons for Churn

As we are investigating what factors cause users to churn, we also want to know what are the most common reasons for users to stop utilizing a particular phone service. Figure 3 shows that the most frequently mentioned reasons come from 2 categories: one is competitors' strength, another one is service providers' customer service quality.

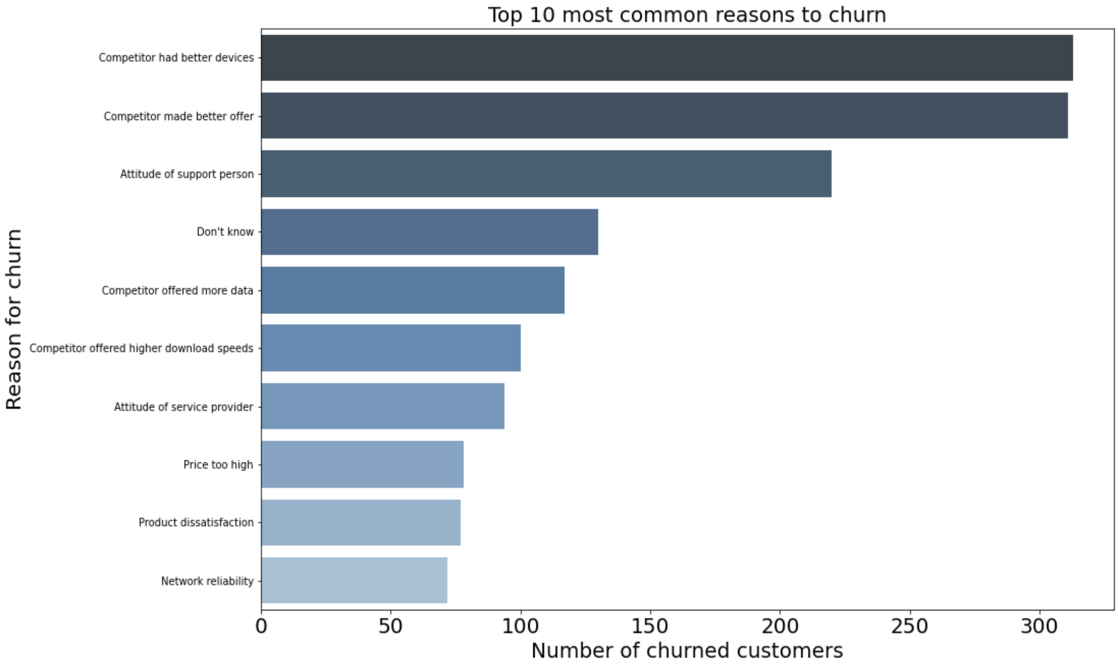


Figure 3. Top 10 most common reasons for churn

d) Tenure

Another assumption we have is that long-term service users are more loyal and less likely to churn even if a competitor offers a better option.

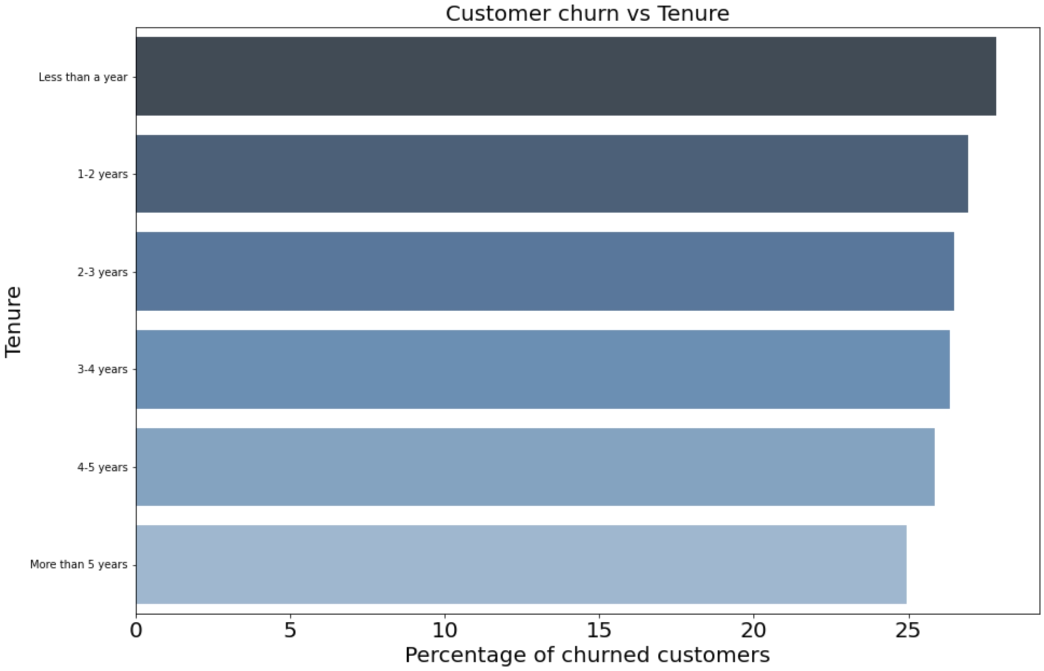


Figure 4. Churned customers rate (%) for different length of tenure

Our assumption is correct, but the difference in percentage for churned customers between those who are in service for less than a year and long-term customers is only 3%.

d) Difference in satisfactions scores for churned/remained customers

Average satisfaction score for churned customers is 1.74, while for remaining customers is 3.79.

e) Discounted offers

A company initialized 5 different discounted offer options. We compare these offers to see what was the most effective and led to the lowest number of churned customers.

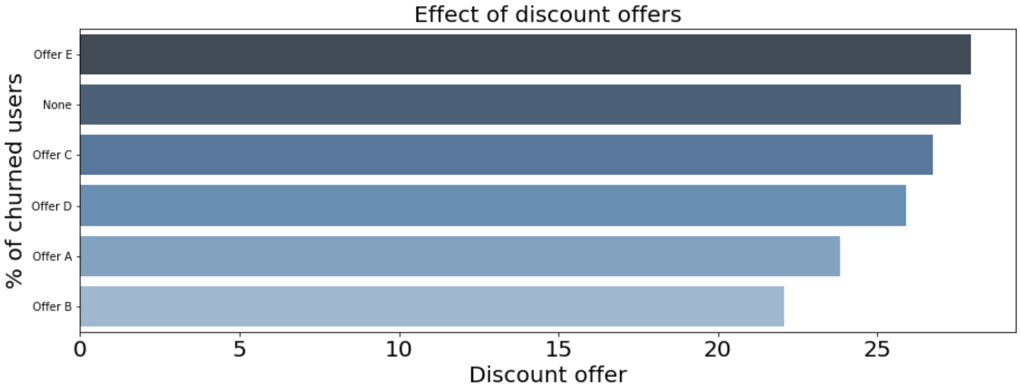


Figure 5. Churned customers rate (%) for different discount offers

As can be seen from Figure 5, Offer E was almost equal to non-discounted offer whereas offer B was the most successful and had 5% less churned customers.

5. Analysis and Interpretation

With the processed dataset, propensity score matching was conducted only excluding the binary outcome "Churn Value" variable while taking all other covariates that may affect whether customers take a discount offer or not (treated or not) into consideration. First, a logistic regression is fitted to predict whether each observation will be in the treatment group. And then, a common support graph, which is the probability distribution of being included in the treatment group for both groups is plotted(Figure 6). It showed large enough overlaps under two distributions. This means our control group and treatment group have very similar distribution of the probability to be treated which is great. After the model fit, a matched sample using k:1 nearest neighbor method was obtained. The weight of each observation was adjusted according to record frequencies. The final matched data contains 5284 records.

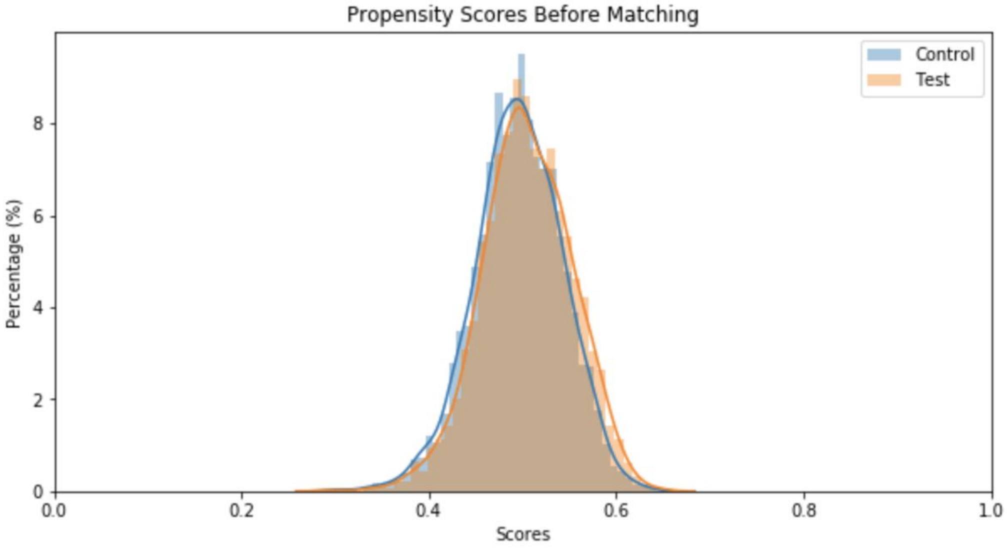


Figure 6. Propensity Scores Before Matching for treatment (test) and control group

Before matching, a logistic regression was fit as an initial check to identify covariates which might have high impact. The model was able to obtain 0.847 AUC. The coefficient for the odds ratio for the treatment variable is 0.1051, suggesting that customers with a phone plan discount offer are less likely to churn compared to those without a phone plan discount offer. However, it should be noted that such a relationship is not statistically significant. It was identified that covariates Dependents, Number_of_Dependents, Referred_a_friend, Number_of_Referrals, Tenure_in_Months, Phone_Service, Online_Security, Online_Backup, Premium Tech Support, Monthly_Charge, and Population might be statistically significant confounders.

After matching, t-tests were conducted to check covariate balance. The p-values for most covariates were not statistically significant which indicates most of the covariates achieved balanced after matching. The three covariates which are statistically significant are: Satisfaction_Score, Internet_Type_Cable, and Internet_Type_DSL. However, as these three covariates were not statistically significant affecting the outcome churning probability, it is reasonable to assume we have achieved covariate balance before proceeding to the next modeling process.

A weighted least-squared model was then fit, including the treatment variable and covariates, using the weight obtained by propensity score matching. From the regression table result, overall the model is obtaining 0.5970 R-squared value. Customers with a phone plan offer appear to have 0.7038 lower probability of churning with statistical significance. The conclusion is in line with our initial assumption.

It can also be seen that Dependents, Paperless_Billing, Total_Refunds, Population, Internet_Service, Internet_Type, and Payment_Method are statistically significant covariates that affect the outcome churning probability. Customers with dependents have 0.0485 less probability of churning with statistical significance. It might be due to the fact that families tend to order the same phone plan with the same service provider and therefore more likely to continue compared to customers without any dependents. Interestingly, customers who opt for paperless billing also have 0.0213 less probability churning compared to those do not. It is likely that they have attached online accounts with the telecompany and might be troublesome to switch. Customers like refunds. It is natural to see that with each unit increase of total refunds, the probability of customer churning decreases by 0.0013. With each unit increase in the current population for that local area, the probability of customer churn decreases by 0.0605. This could be a result of a seller market effect where customers tend to stay as they have limited choices of service providers who have the pricing power. Customers with internet services have 0.1536 higher probability of churning compared to those without. This could be possibly explained as internet service providers are competitors to the tele company providing the same service and there are more choices for customers to choose in terms of internet services and types.

6. Conclusion

Offering customers discounted options decreases the probability of churning by 70% with statistical significance. On top of our matched dataset using propensity scores, this correlation can be extended to causation which fits our initial assumption. In summary, our recommendation for this telecommunication service provider company is to focus on identifying at-risk customers and offering them discount offers to retain the old customers as well as to attract the new ones.

Although we are able to obtain a casual relationship between a discount offer and customer churn based on existing data, there exist other factors that may need to be taken into consideration. These include customers' employment status, education level, number of online or store visits, frequency and means of complaints, number of calls to customers' support, local competitors' moves, etc. These were not taken into account of our analysis but could be valuable insights to understand customer churn.

Furthermore, this analysis cannot be generalized for users located outside of California, so for the future work we suggest collecting more data from different regions. In addition, this data was collected from only one service provider. We suggest comparing different providers between each other, ideally with tracking a user churning from one company to another to compare the difference in plans.

Reference

[1] Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414-1425.

[2] Cognos Analytics - IBM Business Analytics Community. (2020). Retrieved 25 April 2020, from https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

[3] (2020). Retrieved 25 April 2020, from https://data.census.gov/cedsci/table?g=0400000US06&tid=ACSST1Y2018.S1902&t=Income%20and%20Earnings%3AIncome%20and

Appendix (Codebook)

Demographics

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Gender: The customer’s gender: Male, Female

Age: The customer’s current age, in years, at the time the fiscal quarter ended.

Senior Citizen: Indicates if the customer is 65 or older: Yes, No

Married: Indicates if the customer is married: Yes, No

Dependents: Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.

Number of Dependents: Indicates the number of dependents that live with the customer.

Location

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Country: The country of the customer’s primary residence.

State: The state of the customer’s primary residence.

City: The city of the customer’s primary residence.

Zip Code: The zip code of the customer’s primary residence.

Lat Long: The combined latitude and longitude of the customer’s primary residence.

Latitude: The latitude of the customer’s primary residence.

Longitude: The longitude of the customer’s primary residence.

Population

ID: A unique ID that identifies each row.

Zip Code: The zip code of the customer’s primary residence.

Population: A current population estimate for the entire Zip Code area.

Services

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Quarter: The fiscal quarter that the data has been derived from (e.g. Q3).

Referred a Friend: Indicates if the customer has ever referred a friend or family member to this company: Yes, No

Number of Referrals: Indicates the number of referrals to date that the customer has made.

Tenure in Months: Indicates the total amount of months that the customer has been with the company by the end of the quarter specified above.

Offer: Identifies the last marketing offer that the customer accepted, if applicable. Values include None, Offer A, Offer B, Offer C, Offer D, and Offer E.

Phone Service: Indicates if the customer subscribes to home phone service with the company: Yes, No

Avg Monthly Long Distance Charges: Indicates the customer's average long distance charges, calculated to the end of the quarter specified above.

Multiple Lines: Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No

Internet Service: Indicates if the customer subscribes to Internet service with the company: No, DSL, Fiber Optic, Cable.

Avg Monthly GB Download: Indicates the customer's average download volume in gigabytes, calculated to the end of the quarter specified above.

Online Security: Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No

Online Backup: Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No

Device Protection Plan: Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No

Premium Tech Support: Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No

Streaming TV: Indicates if the customer uses their Internet service to stream television programing from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Streaming Movies: Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Streaming Music: Indicates if the customer uses their Internet service to stream music from a third party provider: Yes, No. The company does not charge an additional fee for this service.

Unlimited Data: Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No

Contract: Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.

Paperless Billing: Indicates if the customer has chosen paperless billing: Yes, No

Payment Method: Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check

Monthly Charge: Indicates the customer's current total monthly charge for all their services from the company.

Total Charges: Indicates the customer's total charges, calculated to the end of the quarter specified above.

Total Refunds: Indicates the customer's total refunds, calculated to the end of the quarter specified above.

Total Extra Data Charges: Indicates the customer's total charges for extra data downloads above those specified in their plan, by the end of the quarter specified above.

Total Long Distance Charges: Indicates the customer's total charges for long distance above those specified in their plan, by the end of the quarter specified above.

Status

CustomerID: A unique ID that identifies each customer.

Count: A value used in reporting/dashboarding to sum up the number of customers in a filtered set.

Quarter: The fiscal quarter that the data has been derived from (e.g. Q3).

Satisfaction Score: A customer's overall satisfaction rating of the company from 1 (Very Unsatisfied) to 5 (Very Satisfied).

Satisfaction Score Label: Indicates the text version of the score (1-5) as a text string.

Customer Status: Indicates the status of the customer at the end of the quarter: Churned, Stayed, or Joined

Churn Label: Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.

Churn Value: 1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label.

Churn Score: A value from 0-100 that is calculated using the predictive tool IBM SPSS Modeler. The model incorporates multiple factors known to cause churn. The higher the score, the more likely the customer will churn.

Churn Score Category: A calculation that assigns a Churn Score to one of the following categories: 0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, and 91-100

CLTV: Customer Lifetime Value. A predicted CLTV is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High value customers should be monitored for churn.

CLTV Category: A calculation that assigns a CLTV value to one of the following categories: 2000-2500, 2501-3000, 3001-3500, 3501-4000, 4001-4500, 4501-5000, 5001-5500, 5501-6000, 6001-6500, and 6501-7000.

Churn Category: A high-level category for the customer's reason for churning: Attitude, Competitor, Dissatisfaction, Other, Price. When they leave the company, all customers are asked about their reasons for leaving. Directly related to Churn Reason.

Churn Reason: A customer's specific reason for leaving the company. Directly related to Churn Category.