Survival Analysis of Telco Customers

Ryan Lattanzi

January 21, 2019

Search Results

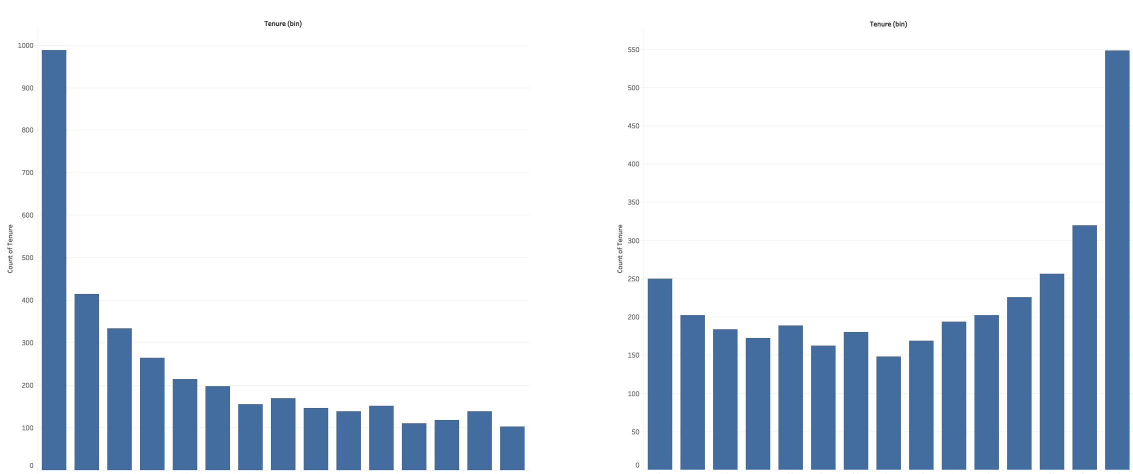
I chose to base my study on the term "customer retention" because customer loyalty is a key element in the longevity of a successful business. This search yielded only two datasets, both of which recorded information about customers including "churn" as the boolean response variable.

Choosing the Dataset

Deciding between the two boiled down to my analysis of the features available. While the `Telecom_customer_attrition` dataset contained information about the types of calls customers made and their charges, it was limited in its lack of variability; I felt this was not a good representation of the customer as a whole. However, the `Telco Customer Churn` dataset painted a better picture of each customer by including demographic information, account information, and services. Hence I chose the latter as a source of investigation.

The Problem

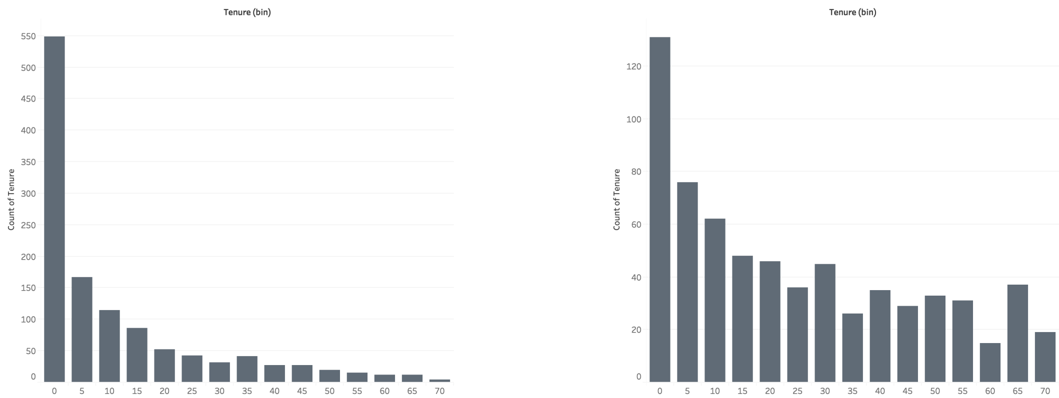
After inputting this data into Tableau for a quick illustration, I noticed a vast difference in the tenure of single customers and customers with a partner shown in the histograms below. Additionally, of the customers who churned, Figure 2 indicates a significant amount of churns occur within the first five months of being with the company. From here I wanted to further investigate other features to see if they contributed to this skewed data. Then using this information, I would construct a model that would attempt to explain the variation in the output.



(a) Customers With No Partner

(b) Customers With Partner

Figure 1: Tenure histograms of customers with or without partners.



(a) Customers With No Partner (b) Customers With Partner

Figure 2: Customers who churn do so within the first five months.

Approach

This problem revealed the perfect opportunity for a survival analysis to estimate probabilities that the customers will stay over time. When contemplating using other kinds of regression models to attempt to predict "survival" times until a customer churned, the following reasons indicated this would not be adequate:

- We could potentially get negative numbers as outputs.
- We need to include censoring for the customers who have not yet left.

In the case of this dataset, the second bullet point is crucial, since only about 26.5% customers actually churned. Going forward, I further peeked into the data via more Tableau illustrations, compared Kaplan - Meier curves, and constructed a Cox regression model in R that provided information about estimated survival probabilities over time.

Results and Discussion

Figure 3 shows notable skewed traits for customers churning within the first five months. The vast majority of these customers have no partner, no dependents, phone service, internet service, no tech support, and month-to-month plans.

Four different Kaplan-Meir curves are depicted in Figure 4, where we compare survival rates of customers with or without partners, if they have multiple lines, whether or not they are senior citizens, and if they are male or female. The p-value is also included on each plot (a p-value smaller than 0.05 indicates a significant difference in survival rates). We can see partners, multiple lines, and seniors result in significant differences while gender does not. It is interesting that the multiple lines feature has a significant p-value while the KM curve does not seem to show a big difference.

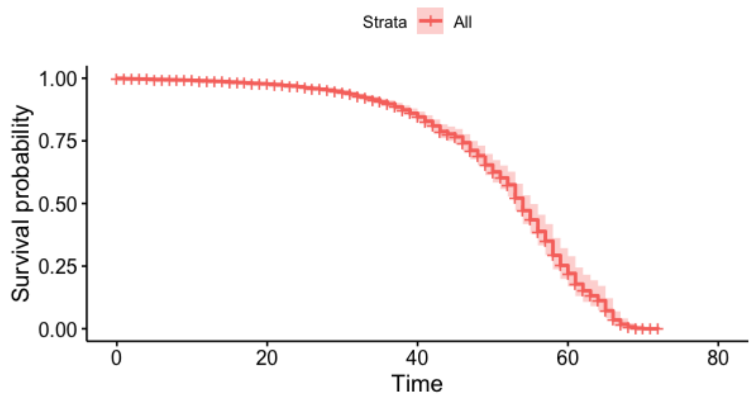
The Cox regression model was first fit using all features. A second model was constructed by backwards elimination so that the resulting features were all significant:

```
Call:
coxph(formula = sur ~ Partner + InternetService + Contract +
    PaperlessBilling + PaymentMethod + TotalCharges, data = telco)

  n= 7043, number of events= 1869
                                     coef  exp(coef)   se(coef)        z Pr(>|z|)
PartnerYes                     -2.128e-01  8.083e-01  5.042e-02   -4.220 2.44e-05 ***
InternetServiceFiber optic      1.780e+00  5.929e+00  6.776e-02   26.268  < 2e-16 ***
InternetServiceNo              -2.022e+00  1.324e-01  1.313e-01  -15.403  < 2e-16 ***
ContractOne year               -1.213e+00  2.973e-01  9.809e-02  -12.368  < 2e-16 ***
```

```
ContractTwo year                       -3.555e+00  2.858e-02  1.968e-01 -18.062  < 2e-16 ***
PaperlessBillingYes                     2.085e-01  1.232e+00  5.565e-02   3.747 0.000179 ***
PaymentMethodCredit card (automatic) -3.036e-02  9.701e-01  9.073e-02  -0.335 0.737902
PaymentMethodElectronic check           4.745e-01  1.607e+00  7.219e-02   6.572 4.95e-11 ***
PaymentMethodMailed check               5.025e-01  1.653e+00  8.666e-02   5.798 6.70e-09 ***
TotalCharges                           -1.353e-03  9.986e-01  3.312e-05 -40.841  < 2e-16 ***
```

Notice the significant predictors are similar to those seen in Figure 3. However the dependents, phone service, and tech support features are not included because they are highly correlated to the partner and internet service features which is easily seen in Figure 3. It is also interesting to note that the predictors in our final model do not contain features that we might have expected. For example, the KM curve for the senior citizen feature shows a significant change in survival rates, but it is not included in our model. This can easily be explained by the model measuring significance when considering all features together, rather than independently. A plot of estimated survival probabilities using the final model is shown below.



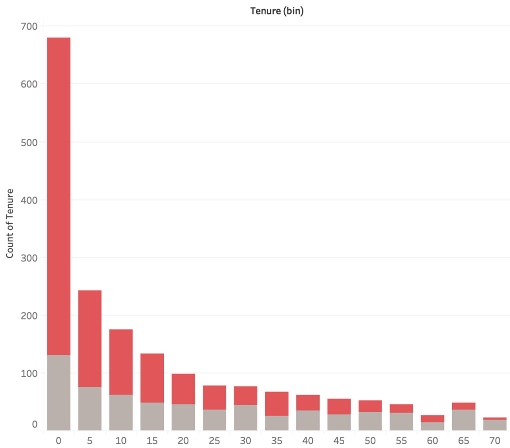
Further Investigations

First off I want to note that the dataset, while containing great variables, was still missing a couple that I think would be important. A huge component to include would be a categorical "reason for leaving" variable as this cannot be inferred from the data now. Then we would have a better idea of what to improve on to potentially retain those customers. We could contact these individuals and get this data if we have the resources and feel it is important enough. Also, I think the data should include whether or not a customer has returned after churning in order to know if the customer was retained in the long run.

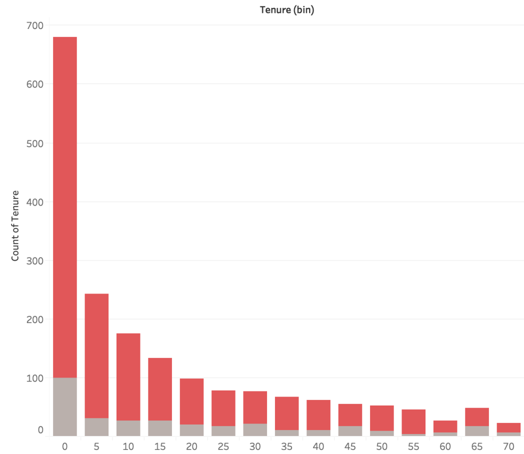
As for what we have discovered in the previous section, it is clear the main group to target consists of singles with month-to-month plans. Perhaps incentives could be provided for customers on a monthly plan if they stay longer than x number of months. Or, we could improve on longer contract deals so that customers would prefer them over a monthly basis. But I will leave the particulars up to the marketing and sales professionals.

Moving forward, it may be worthwhile to look into the following:

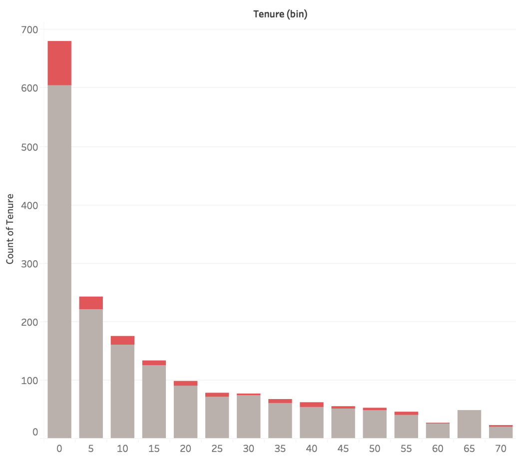
- Why is it that customers with partners have a lower churn rate? (Singles make up 64% of churned customers)
- We can see the survival probability begins a more drastic dip after about 24 months. What can we do to prevent this and keep customers longer?
- In what ways can we diversify our services to attract and retain more customers?



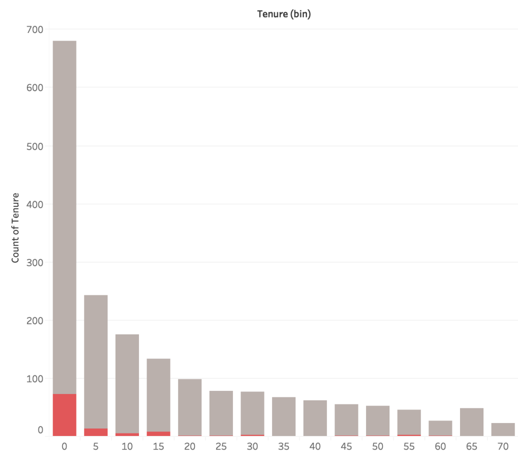
(a) Customers With No Partner (red)



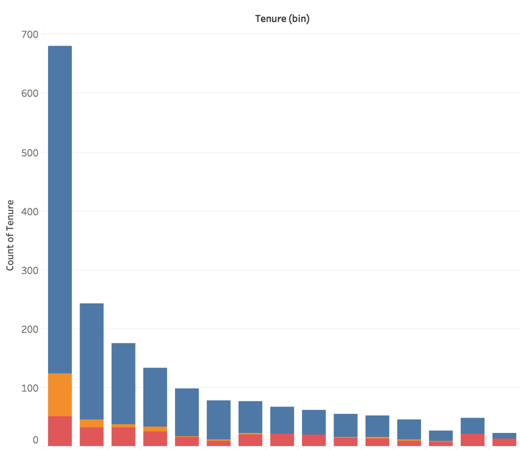
(b) Customers With No Dependents (red)



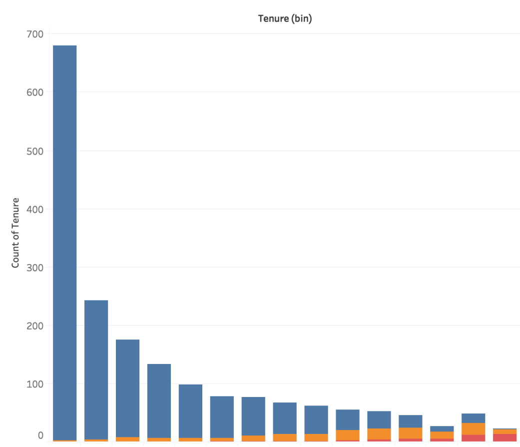
(c) Customers With Phone Service (gray)



(d) Customers With Internet (gray)

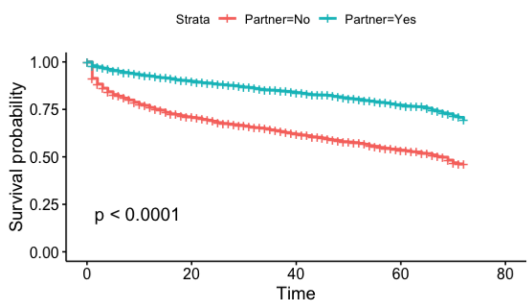


(e) Customers Without Tech Support (blue)

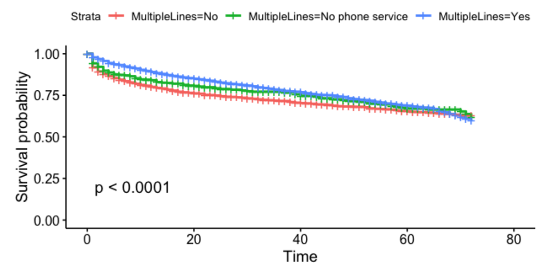


(f) Month-to-Month Customers (blue)

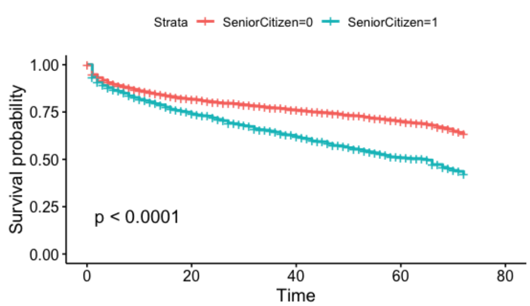
Figure 3: Notable traits for churning customers.



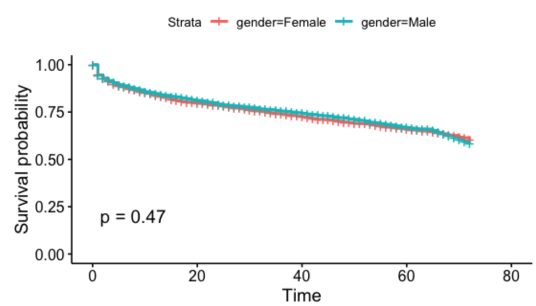
(a) Partners



(b) Multiple Lines



(c) Senior



(d) Gender

Figure 4: Notable traits for churning customers.