Iranian Churn Data Exploratory Analysis

Tim Pulfer, Jacob Rosen, Dhun Sheth

Statistical Description of Data

The dataset comprises a comprehensive overview of customer interactions and characteristics within a telecommunications context over a 9-month period, focusing on understanding customer churn. It consists of 13 variables, which are a mix of numerical, binary, and ordinal types. These variables capture a wide range of information, including customer usage patterns, service satisfaction, subscription details, demographic profiles, and overall customer value. This structured data collection facilitates a multifaceted analysis of factors influencing customer retention and turnover, allowing for targeted insights into service improvements and customer engagement strategies. The database has a total of 3150 rows, each representing a customer. The following are the 13 columns in the database:

- Call Failures (numerical): the number of call failures over 9 month period
- Complains (binary): 0: No complaint, 1: complaint
- Subscription Length (numerical): total months of subscription
- Charge Amount (ordinal): 0: lowest amount, 9: highest amount
- Seconds of Use (numerical): total seconds of calls over 9 month period
- Frequency of use (numerical): total number of calls over 9 month period
- Frequency of SMS (numerical): total number of text messages over 9 month period
- Distinct Called Numbers (numerical): total number of distinct phone calls
- Age Group (ordinal): 1: younger age, 5: older age (10-19 is 1, 20-29 is 2, etc.)
- Age (numerical): age of customer (24, 25, 30, etc.)
- Tariff Plan (binary): 1: Pay as you go, 2: contractual
- Status (binary): 1: active, 2: non-active
- Customer Value (numerical): The calculated value of customer 9 month period
- Churn (binary): 1: churn, 0: non-churn - the state of the customers at the end of 12 months

Table 1: Summary statistics of all predictors

Variable	Mean	Std	Min	25%	50%	75%	Max
Call Failure	7.63	7.26	0.00	1.00	6.00	12.00	36.00
Complains	0.08	0.27	0.00	0.00	0.00	0.00	1.00
Subscription Length	32.54	8.57	3.00	30.00	35.00	38.00	47.00
Charge Amount	0.94	1.52	0.00	0.00	0.00	1.00	10.00
Seconds of Use	4472.46	4197.91	0.00	1391.25	2990.00	6478.25	17090.00
Frequency of use	69.46	57.41	0.00	27.00	54.00	95.00	255.00
Frequency of SMS	73.17	112.24	0.00	6.00	21.00	87.00	522.00
Distinct Called Numbers	23.51	17.22	0.00	10.00	21.00	34.00	97.00
Age Group	2.83	0.89	1.00	2.00	3.00	3.00	5.00
Tariff Plan (Pay On Contract)	0.08	0.27	0.00	0.00	0.00	0.00	1.00
Status (Inactivity)	0.25	0.43	0.00	0.00	0.00	0.00	1.00
Age	30.99	8.83	15.00	25.00	30.00	30.00	55.00
Customer Value	470.97	517.02	0.00	113.80	228.48	788.39	2165.28
Churn	0.16	0.36	0.00	0.00	0.00	0.00	1.00

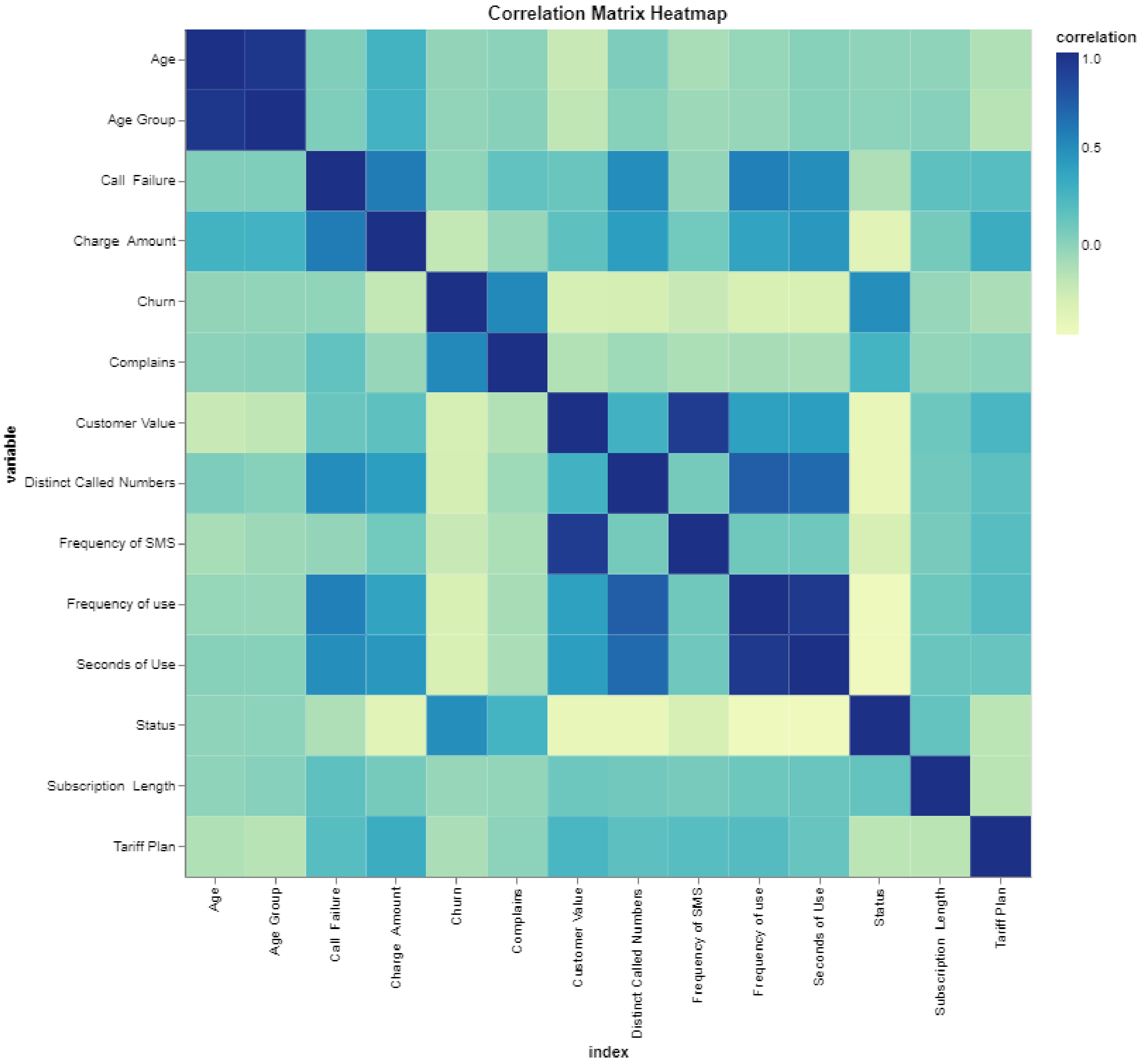


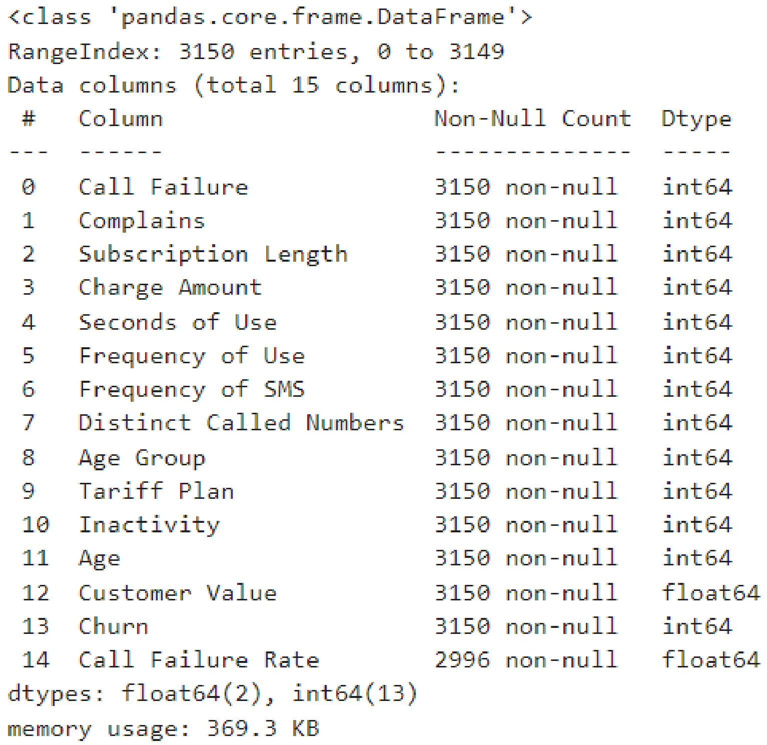
Figure 1: Correlation matrix for predictors

Data Transformations

1. The “Status” column was changed from 1: Active, 2: Non-Active to 0: Active, 1: In-active - to match the churn column where 0: non-churn, 1: churned.
2. The ‘Tariff Plan’ column was changed from 1: Pay as you go, 2: Contractual to 0: False, 1: True.
3. The “Status” column was renamed to “Inactivity” because this title better represents the column.
4. The ‘Tariff Plan’ column was renamed to ‘Pay On Contract’ to match the binary values.
5. Calculated a new field based on # of Call Failures and Frequency of Use, called “Call Failure Rate”.

Null

1. Set the null values in the Call Failure Rate column to the mean of the column.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3150 entries, 0 to 3149
Data columns (total 15 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Call Failure             3150 non-null   int64  
 1   Complains                3150 non-null   int64  
 2   Subscription Length      3150 non-null   int64  
 3   Charge Amount            3150 non-null   int64  
 4   Seconds of Use           3150 non-null   int64  
 5   Frequency of Use         3150 non-null   int64  
 6   Frequency of SMS         3150 non-null   int64  
 7   Distinct Called Numbers  3150 non-null   int64  
 8   Age Group                3150 non-null   int64  
 9   Tariff Plan              3150 non-null   int64  
 10  Inactivity               3150 non-null   int64  
 11  Age                      3150 non-null   int64  
 12  Customer Value           3150 non-null   float64
 13  Churn                    3150 non-null   int64  
 14  Call Failure Rate        2996 non-null   float64
dtypes: float64(2), int64(13)
memory usage: 369.3 KB
```

Figure 2: Null Values in Dataset

The only column with null values was the calculated "Call Failure Rate" column which had nulls as a result of the 0/0 division. Rather than setting the null values to 0, they were changed to the mean value of the column. They were not set to 0 because that would have added more weight to the rows where it was 0, due to the # of call failures being 0 but the frequency being greater than 0.

Outliers

Cook's distance was considered to detect outliers, however, this metric measures the effect the deletion of an observation has on a linear regression line. When considering the complex relationships and various modeling options, it could be likely the relationship between the response and a column will be non-linear and so deleting an observation based on the effect it has for a linear relationship is not sensible. In addition, although outliers based on the inter-quantile range were detected, it was hypothesized that these extreme values may hold some predictive information as to customers churning or not. As such, for the outlier analysis, no outliers were removed, however, values were analyzed to ensure they made sense and illogical values would be removed. No illogical values were observed and so no observations were removed.

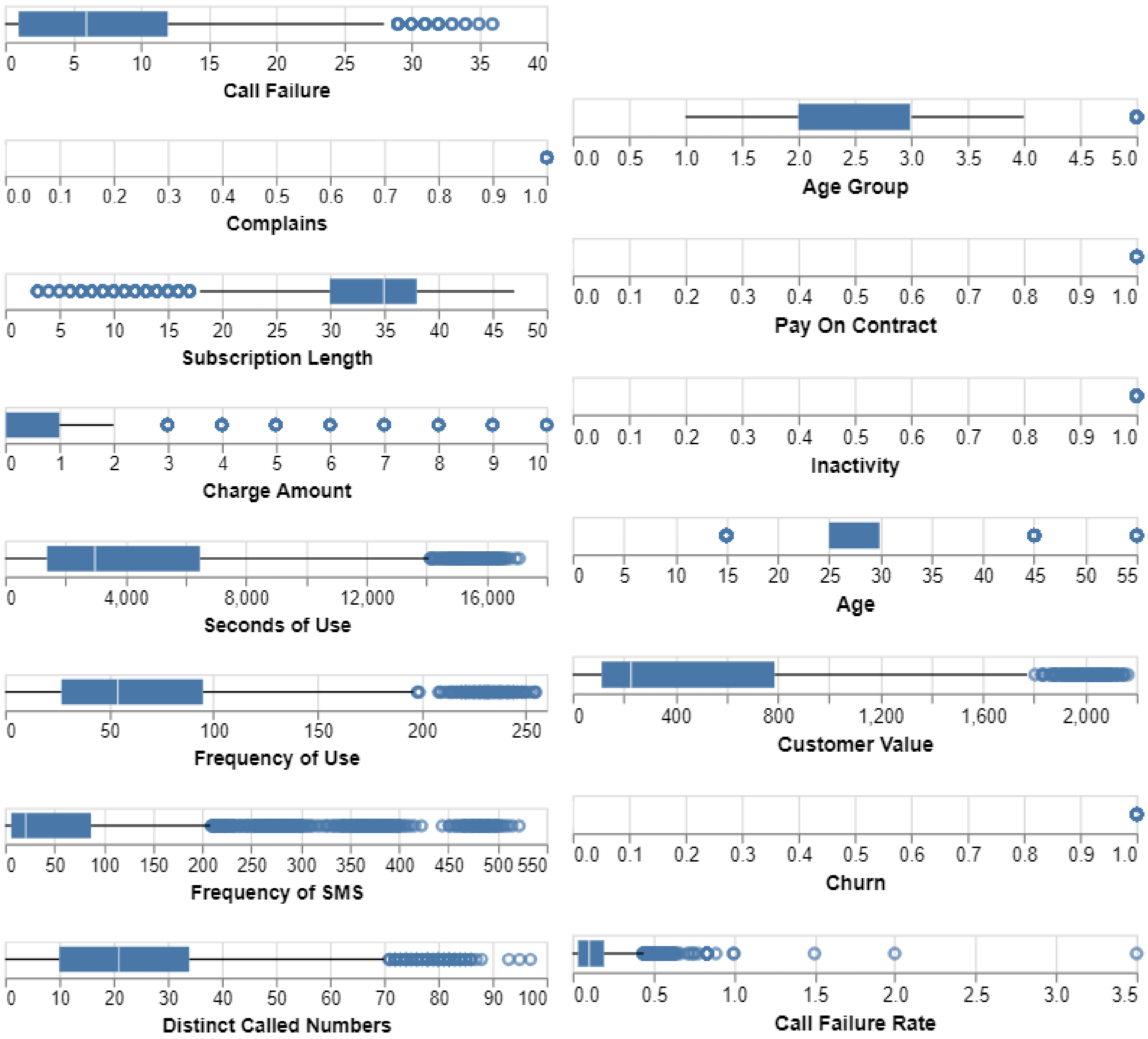


Figure 3: Boxplots of each column in the dataset

Balance of Churn

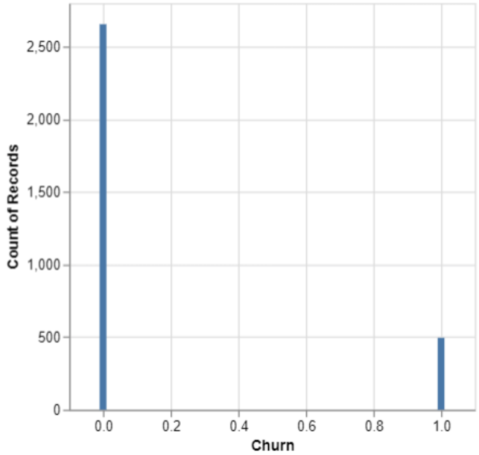


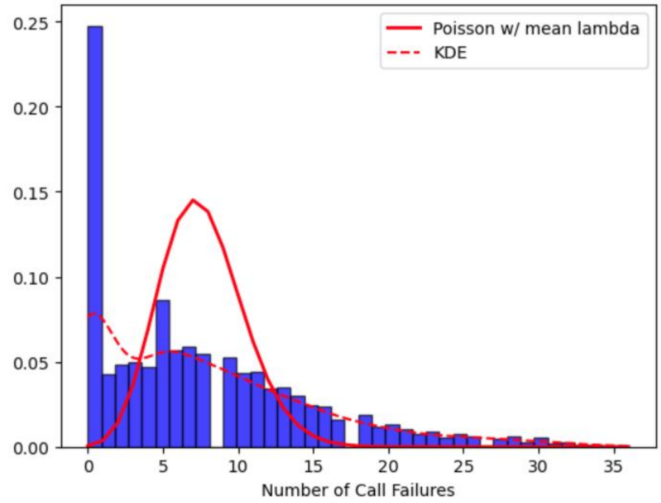
Figure 4: Count of Churn, 0: Not Churned, 1: Churned

As seen above, the data is very unbalanced with 500 rows of churned customers, and the remaining not churning. Although imbalance is not always an issue, it is critical to understand this is occurring and is accounted for. As an example, classification accuracy or misclassification rates tend to be very deceiving for unbalanced groups, like in this example, so looking more closely at the confusion matrix to verify this metric becomes more important.

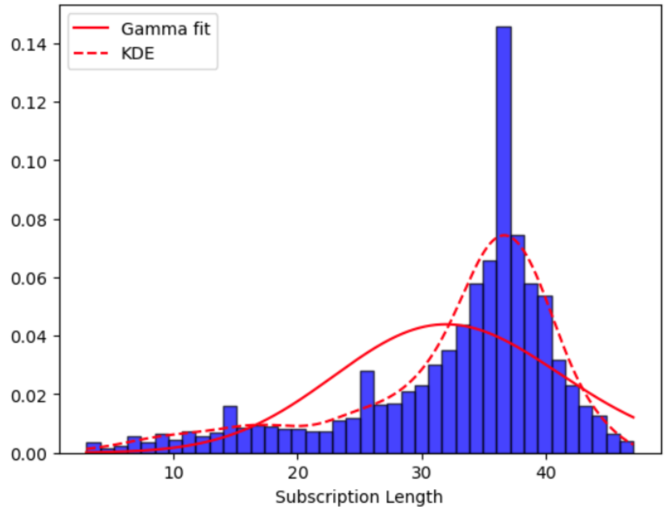
Predictor Distributions

Some columns were skipped as they contained binary or categorical values - only numerical columns were fit.

1. Number of call failures

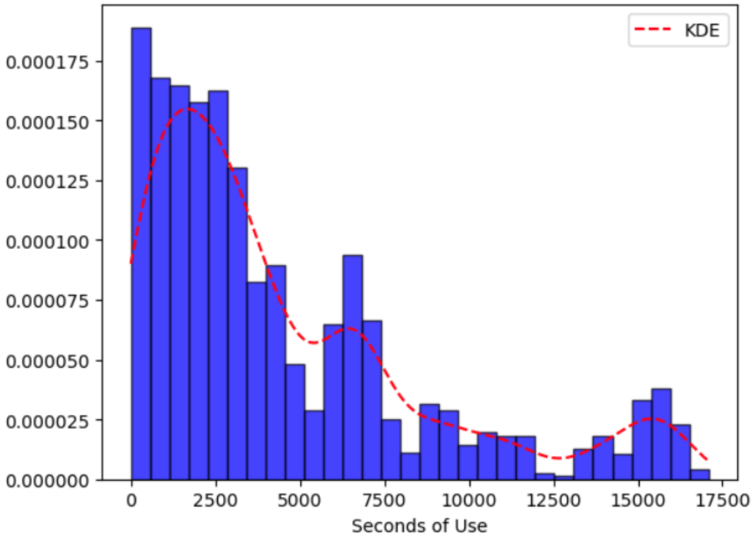


2. Subscription Length

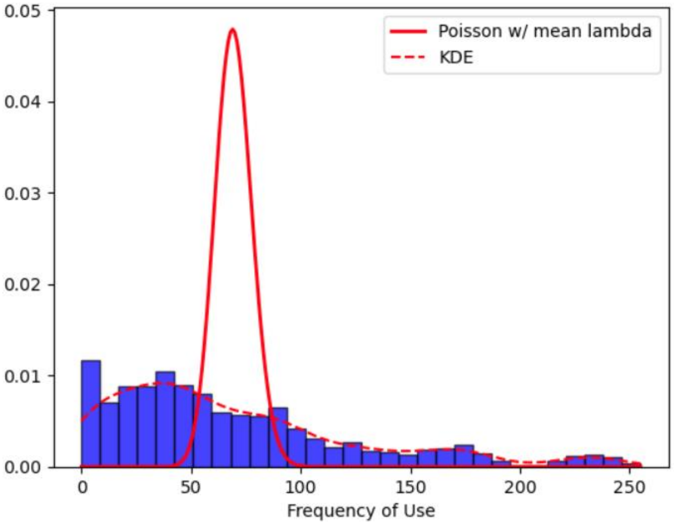


3. Seconds of Use

   Attempted fitting a Gamma and Weibull distribution, but both fits were poor.

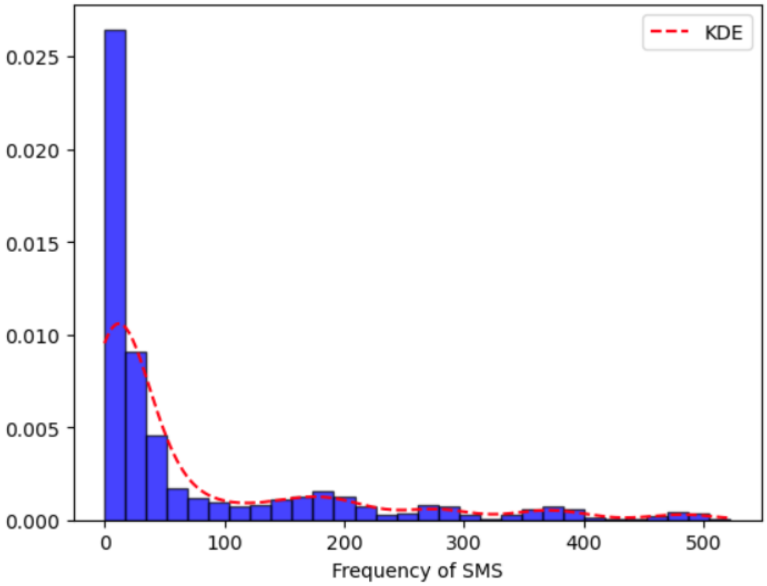


4. Frequency of Use

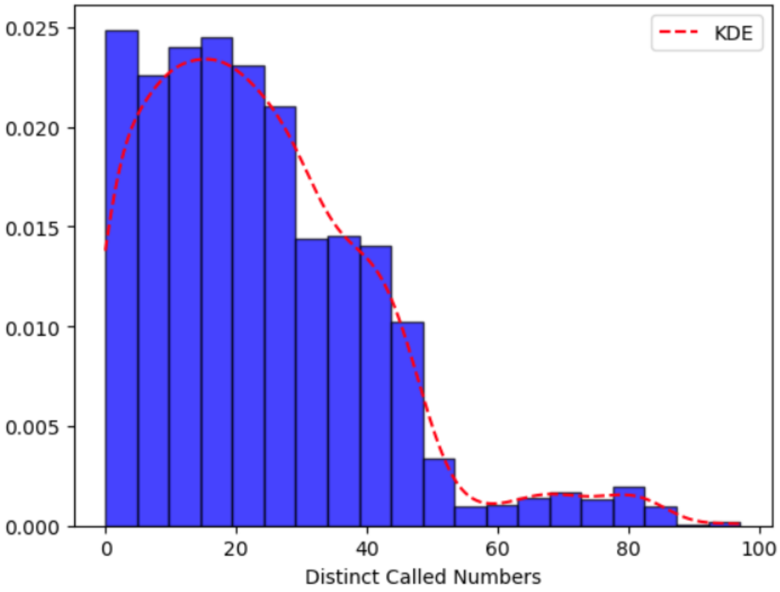


Attempted fitting a Poisson distribution, but the fit looks very bad, as seen from the plot above. The KDE fit much closer to the distribution seen from the histogram. The issue with fitting a Poisson distribution also occurred with several other columns, for which a KDE was fit instead.

5. Number of SMS

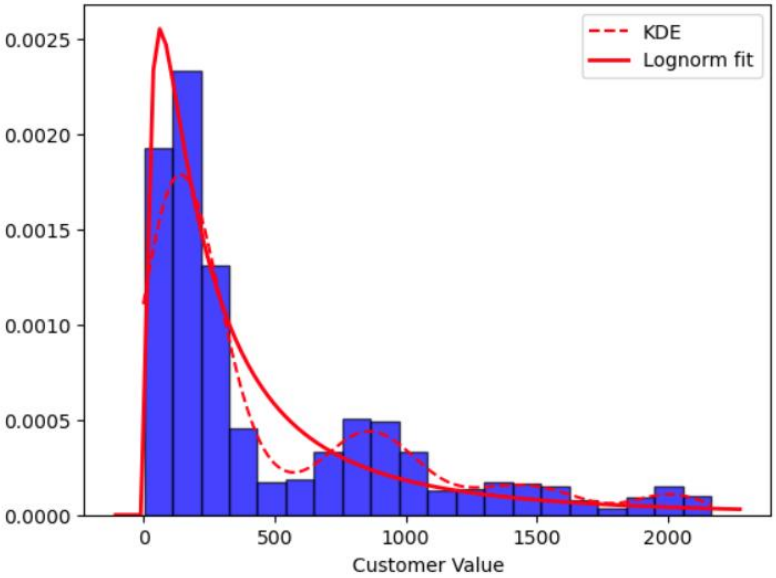


6. Distinct Called Numbers



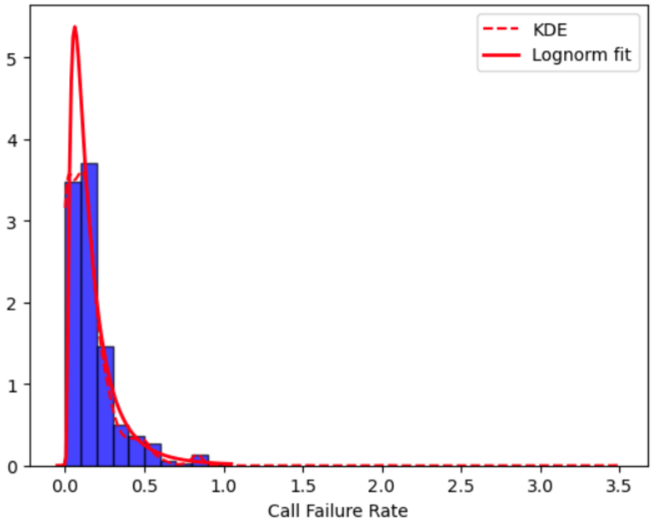
Attempted Poisson but due fit was poor and when plotted over the histogram, due to the y-range plot did not match at all - similar to the Frequency of Use plot.

7. Customer Value

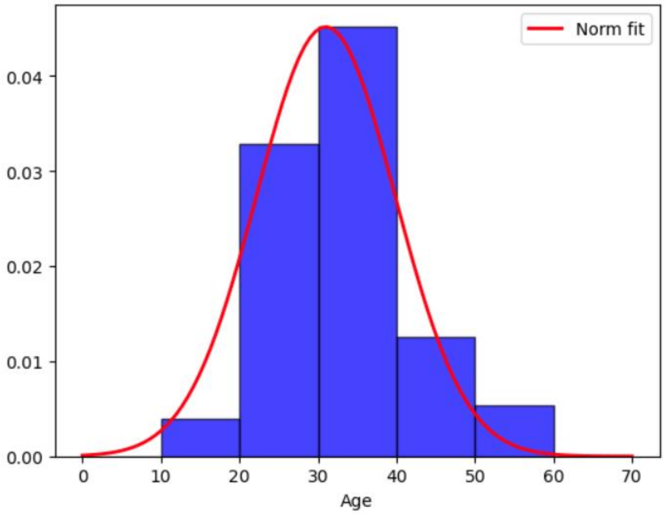


Had to remove 0 values for lognorm fit, hence KDE may be better because it fits better and 0-values don't need to be excluded.

8. Call Failure Rate



9. Age



Scientific Questions

Our project aims to explore the following scientific questions with regard to the customer churn dataset:

- Which variables are most important for predicting customer churn?

To determine the most important predictors for churn we can calculate the marginal distributions of each feature in each group and use a z-test to check the significance of the difference between the distributions. We can also build classification models and check the significance of each variable in the models.

- What is the call failure rate at which we expect a customer to churn?

We can determine this by creating a sample from our fitted distributions of each variable and, using the model of churn we created in the previous question, predict the churn on our simulated data. With the predictions, we can analyze the call failure rates from the predicted churn observations to build a confidence interval for the call failure rate at which we expect a customer to churn.

- What is the number of call failures before we can expect a complaint?

Similar to our previous question, we can model the relationship between call failures and complaints and use it for predictions on simulated data. Again, we can build confidence intervals to determine the number of call failures that result in a complaint.

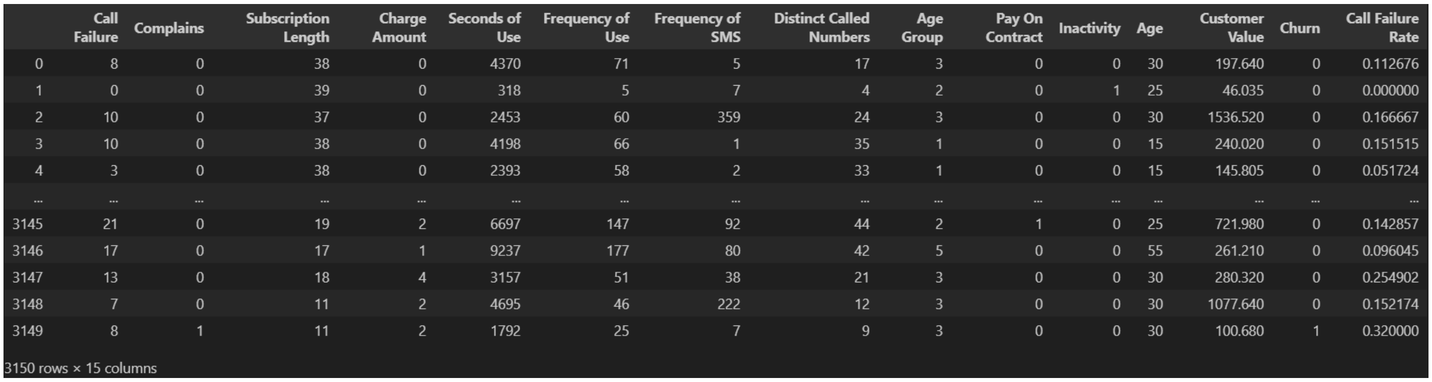
- Are these thresholds the same for high-valued customers as they are for low-valued customers?

We should be able to determine this by fitting a full model. If the customer value variable is significant it means there is evidence to show that high-valued and low-valued customers have different churn patterns based on the other variables in the model. If we want to inspect this further, we can bin the observations into high and low-valued customers and use the binned variable in the model to determine a coefficient for each group.

- What is the expected reduction in churn if a proposed business solution can reduce call failures by 50%?

We can determine this by shifting our fitted distribution for call failures to reflect a reduction by 50% and create a new simulated sample. We can then use our model to predict churn on the new sample and compare it to results from our previous sample to determine the reduction in churn.

Dataset after transformations:



	Call Failure	Complains	Subscription Length	Charge Amount	Seconds of Use	Frequency of Use	Frequency of SMS	Distinct Called Numbers	Age Group	Pay On Contract	Inactivity	Age	Customer Value	Churn	Call Failure Rate
0	8	0	38	0	4370	71	5	17	3	0	0	30	197.640	0	0.112676
1	0	0	39	0	318	5	7	4	2	0	1	25	46.035	0	0.000000
2	10	0	37	0	2453	60	359	24	3	0	0	30	1536.520	0	0.166667
3	10	0	38	0	4198	66	1	35	1	0	0	15	240.020	0	0.151515
4	3	0	38	0	2393	58	2	33	1	0	0	15	145.805	0	0.051724
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3145	21	0	19	2	6697	147	92	44	2	1	0	25	721.980	0	0.142857
3146	17	0	17	1	9237	177	80	42	5	0	0	55	261.210	0	0.096045
3147	13	0	18	4	3157	51	38	21	3	0	0	30	280.320	0	0.254902
3148	7	0	11	2	4695	46	222	12	3	0	0	30	1077.640	0	0.152174
3149	8	1	11	2	1792	25	7	9	3	0	0	30	100.680	1	0.320000

3150 rows × 15 columns