Final of Project Report

Telecom Churn Prediction

Group 20

Manoghn Kandiraju

Saathvika Kethineni

kandiraju.m@northeastern.edu
kethineni.s@northeastern.edu

Percentage of Effort Contributed by Student 1:_____50%_____

Percentage of Effort Contributed by Student 2:_____50%_____

Signature of Student 1:_____M.K._____

Signature of Student 2:_____S.K._____

Submission Date:_____22nd March 2024_____

Problem Setting

This project is situated in the dynamic and competitive landscape of the telecommunications industry, where understanding and managing customer churn – the loss of clients or customers – is critical for sustaining business growth and profitability. With the advancements in data collection and analytics, the project revolves around leveraging a comprehensive dataset to glean insights into customer behavior, preferences, and patterns of service usage.

The key challenge lies in interpreting a complex array of data, encompassing various aspects of customer interactions – from basic demographics to detailed service usage and billing histories. This analysis is pivotal for formulating strategic decisions aimed at enhancing customer retention, personalizing services, and improving overall customer satisfaction.

Problem Definition:

The objective of analyzing the telecommunications dataset is to ascertain the primary factors influencing customer churn, which is the tendency of customers to discontinue their services with the company. This necessitates a rigorous data cleaning process to eliminate any inaccuracies, inconsistencies, or missing values that could potentially skew the analysis results.

Once the dataset is refined, a range of machine learning models will be employed to dissect and understand the intricate relationships between various customer-related variables and their churn behavior. The goal is to pinpoint specific factors — such as service usage patterns, billing information, customer demographics, and service satisfaction levels — that most significantly sway a customer's decision to continue or terminate their services with the company.

To achieve this, diverse machine learning algorithms will be applied to model and predict churn behaviors based on the dataset. These models will undergo a thorough evaluation process using key performance metrics like accuracy, precision, recall, and F1 score, enabling us to gauge their predictive efficacy.

After a comprehensive training and evaluation phase, the most effective model will be chosen for deeper analysis. This model will serve as a tool to identify the most impactful variables in customer churn and will be instrumental in forecasting future trends in customer retention.

In essence, through meticulous analysis of the dataset and application of machine learning models, valuable insights can be gleaned regarding the determinants of customer retention. These insights will empower the company to make informed, data-driven decisions aimed at enhancing customer satisfaction, bolstering motivation, and reducing churn rates. The ultimate goal is to foster a more engaging and satisfying customer experience, thereby decreasing turnover and bolstering the company's overall market performance.

Data Sources:

Dataset: 😵‍💫 Telecom Customer Churn Prediction (kaggle.com)

Data Description:

The dataset comprises **7043** rows and **38** columns, with key variables including:

CustomerID: Unique identifier for each customer.

Gender: Male or Female.

Age: Age of the customer.

Married: Binary variable indicating if the customer is married (Yes) or not (No).

Number of Dependents: Number of dependents of the customer.

City: City of the customer.

Zip Code: Zip code of the customer.

Latitude and Longitude: Geographical coordinates of the customer.

Number of Referrals: Number of customers referred by the customer.

Tenure in Months: Number of months the customer has been with the telecom service.

Offer: Type of offer the customer is subscribed to.

Phone Service: Binary variable indicating if the customer has phone service (Yes) or not (No).

Avg Monthly Long Distance Charges: Average monthly long-distance charges incurred by the customer.

Multiple Lines: Categorical variable indicating if the customer has multiple phone lines.

Internet Service: Type of internet service subscribed by the customer.

Internet Type: Type of internet connection.

Avg Monthly GB Download: Average monthly gigabytes downloaded by the customer.

Online Security, Online Backup, Device Protection Plan, Premium Tech Support, Streaming TV, Streaming Movies, Streaming Music, Unlimited Data: Categorical variables indicating if the customer has these services.

Contract: Type of contract the customer is subscribed to.

Paperless Billing: Binary variable indicating if the customer has paperless billing (Yes) or not (No).

Payment Method: Payment method chosen by the customer.

Monthly Charge: Monthly charges incurred by the customer.

Total Charges: Cumulative charges incurred by the customer.

Total Refunds, Total Extra Data Charges, Total Long Distance Charges, Total Revenue: Additional financial information.

Customer Status: Current status of the customer (Stayed, Joined, Churned).

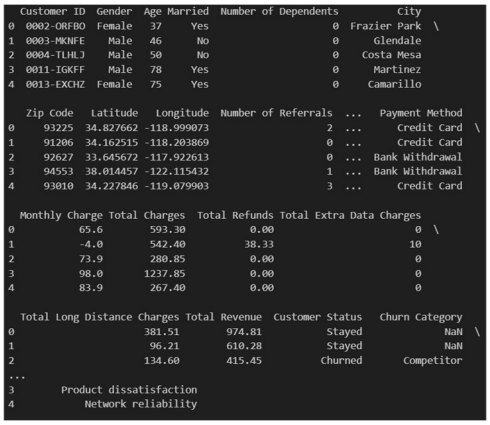
Churn Category: Reason for churn (if applicable).

Churn Reason: Detailed reason for churn (if applicable).

This comprehensive dataset enables a thorough analysis to uncover insights that can drive effective customer retention strategies in the telecom industry.

Data Preprocessing:

1. Loading the dataset and displaying rows:



```
  Customer ID  Gender  Age Married  Number of Dependents          City
0  0002-ORFBO  Female   37     Yes                     0  Frazier Park  \
1  0003-MKNFE    Male   46      No                     0     Glendale
2  0004-TLHLJ    Male   50      No                     0   Costa Mesa
3  0011-IGKFF    Male   78     Yes                     0     Martinez
4  0013-EXCHZ  Female   75     Yes                     0    Camarillo

   Zip Code   Latitude   Longitude  Number of Referrals  ...   Payment Method
0     93225  34.827662 -118.999073                    2  ...      Credit Card  \
1     91206  34.162515 -118.203869                    0  ...      Credit Card
2     92627  33.645672 -117.922613                    0  ...  Bank Withdrawal
3     94553  38.014457 -122.115432                    1  ...  Bank Withdrawal
4     93010  34.227846 -119.079903                    3  ...      Credit Card

  Monthly Charge Total Charges  Total Refunds Total Extra Data Charges
0           65.6         593.30          0.00                       0  \
1           -4.0         542.40         38.33                      10
2           73.9         280.85          0.00                       0
3           98.0        1237.85          0.00                       0
4           83.9         267.40          0.00                       0

  Total Long Distance Charges Total Revenue  Customer Status   Churn Category
0                      381.51        974.81          Stayed              NaN  \
1                       96.21        610.28          Stayed              NaN
2                      134.60        415.45         Churned       Competitor
...
3       Product dissatisfaction
4           Network reliability
```

2. Generating summary statistics of the numerical columns, such as count, mean, standard deviation, minimum, maximum, and quartiles. To get insights into the distribution and central tendency of numerical data.

	Age	Number of Dependents	Zip Code	Latitude	Longitude	Number of Referrals	Tenure in Months	Avg Monthly Long Distance Charges	Avg Monthly GB Download	Monthly Charge	Total Charges	Total Refunds	Total Extra Data Charges	Total Long Distance Charges	Total Revenue
count	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	6361.000000	5517.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000	7043.000000
mean	46.509726	0.468692	93486.070567	36.197455	-119.756684	1.951867	32.386767	25.420517	26.189958	63.596131	2280.381264	1.962182	6.860713	749.099262	3034.379056
std	16.750352	0.962802	1856.767505	2.468929	2.154425	3.001199	24.542061	14.200374	19.586585	31.204743	2266.220462	7.902614	25.104978	846.660055	2865.204542
min	19.000000	0.000000	90001.000000	32.555828	-124.301372	0.000000	1.000000	1.010000	2.000000	-10.000000	18.800000	0.000000	0.000000	0.000000	21.360000
25%	32.000000	0.000000	92101.000000	33.990646	-121.788090	0.000000	9.000000	13.050000	13.000000	30.400000	400.150000	0.000000	0.000000	70.545000	605.610000
50%	46.000000	0.000000	93518.000000	36.205465	-119.595293	0.000000	29.000000	25.690000	21.000000	70.050000	1394.550000	0.000000	0.000000	401.440000	2108.640000
75%	60.000000	0.000000	95329.000000	38.161321	-117.969795	3.000000	55.000000	37.680000	30.000000	89.750000	3786.600000	0.000000	0.000000	1191.100000	4801.145000
max	80.000000	9.000000	96150.000000	41.962127	-114.192901	11.000000	72.000000	49.990000	85.000000	118.750000	8684.800000	49.790000	150.000000	3564.720000	11979.340000

Key statistics include an average age of 46.51 years, with 46.51% having dependents, and a mean monthly charge of $63.60, with standard deviation of $31.20.

3. Deriving the data types of each column, the number of non-null values, and memory usage to identify missing values.

```
#   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Customer ID                       7043 non-null   object
 1   Gender                            7043 non-null   object
 2   Age                               7043 non-null   int64
 3   Married                           7043 non-null   object
 4   Number of Dependents              7043 non-null   int64
 5   City                              7043 non-null   object
 6   Zip Code                          7043 non-null   int64
 7   Latitude                          7043 non-null   float64
 8   Longitude                         7043 non-null   float64
 9   Number of Referrals               7043 non-null   int64
 10  Tenure in Months                  7043 non-null   int64
 11  Offer                             3166 non-null   object
 12  Phone Service                     7043 non-null   object
 13  Avg Monthly Long Distance Charges 6361 non-null   float64
 14  Multiple Lines                    6361 non-null   object
 15  Internet Service                  7043 non-null   object
 16  Internet Type                     5517 non-null   object
 17  Avg Monthly GB Download           5517 non-null   float64
 18  Online Security                   5517 non-null   object
 19  Online Backup                     5517 non-null   object
...
 36  Churn Category                    1869 non-null   object
 37  Churn Reason                      1869 non-null   object
dtypes: float64(9), int64(6), object(23)
```

4. Removing irrelevant features ('Customer ID', 'Zip Code', 'Latitude', 'Longitude', 'City', 'Zip Code')

```
Gender                                0
Age                                   0
Married                               0
Number of Dependents                  0
Number of Referrals                   0
Tenure in Months                      0
Offer                              3877
Phone Service                         0
Avg Monthly Long Distance Charges   682
Multiple Lines                      682
Internet Service                      0
Internet Type                      1526
Avg Monthly GB Download            1526
Online Security                    1526
Online Backup                      1526
Device Protection Plan             1526
Premium Tech Support               1526
Streaming TV                       1526
Streaming Movies                   1526
Streaming Music                    1526
Unlimited Data                     1526
Contract                              0
Paperless Billing                     0
Payment Method                        0
Monthly Charge                        0
...
Total Revenue                         0
Customer Status                       0
Churn Category                     5174
Churn Reason                       5174
```

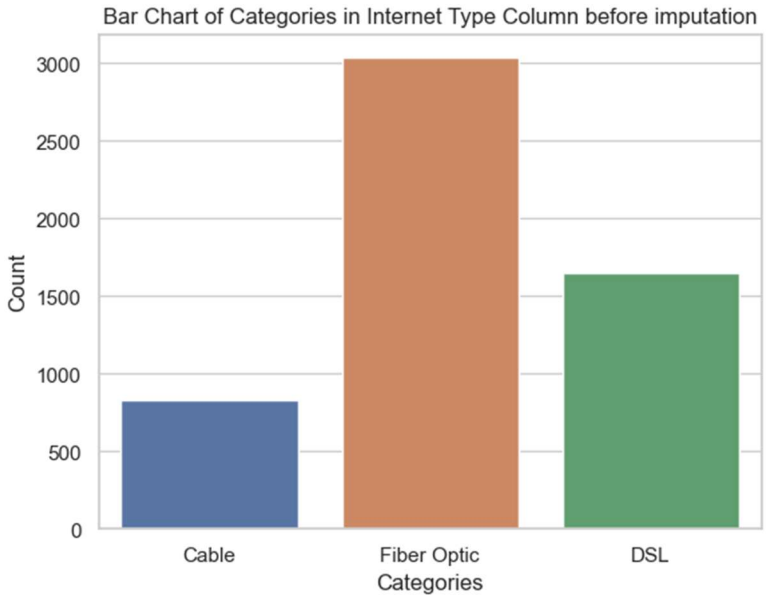
5. Handling missing values by dropping the columns 'Churn Category' and 'Churn Reason'

```
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          7043 non-null   object
 1   Age                             7043 non-null   int64
 2   Married                         7043 non-null   object
 3   Number of Dependents            7043 non-null   int64
 4   Number of Referrals             7043 non-null   int64
 5   Tenure in Months                7043 non-null   int64
 6   Offer                           3166 non-null   object
 7   Phone Service                   7043 non-null   object
 8   Avg Monthly Long Distance Charges  6361 non-null   float64
 9   Multiple Lines                  6361 non-null   object
 10  Internet Service                7043 non-null   object
 11  Internet Type                   5517 non-null   object
 12  Avg Monthly GB Download         5517 non-null   float64
 13  Online Security                 5517 non-null   object
 14  Online Backup                   5517 non-null   object
 15  Device Protection Plan          5517 non-null   object
 16  Premium Tech Support            5517 non-null   object
 17  Streaming TV                    5517 non-null   object
 18  Streaming Movies                5517 non-null   object
 19  Streaming Music                 5517 non-null   object
...
 29  Total Revenue                   7043 non-null   float64
 30  Customer Status                 7043 non-null   object
```

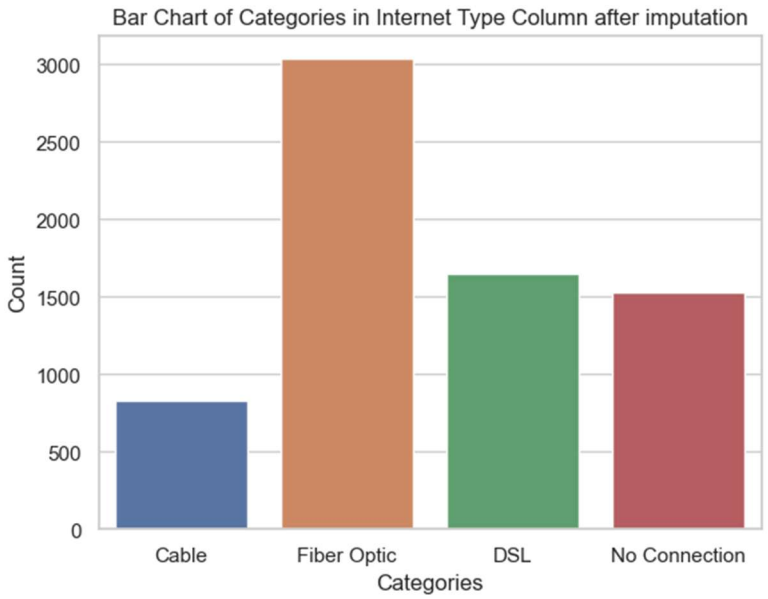
6. Filling missing values in columns 'Avg Monthly Long-Distance Charges' and 'Multiple Lines' with 0 and 'No' respectively, where 'Phone Service' was 'No'.

```
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Gender                             7043 non-null   object
 1   Age                                7043 non-null   int64
 2   Married                            7043 non-null   object
 3   Number of Dependents               7043 non-null   int64
 4   Number of Referrals                7043 non-null   int64
 5   Tenure in Months                   7043 non-null   int64
 6   Offer                              3166 non-null   object
 7   Phone Service                      7043 non-null   object
 8   Avg Monthly Long Distance Charges  7043 non-null   float64
 9   Multiple Lines                     7043 non-null   object
 10  Internet Service                   7043 non-null   object
 11  Internet Type                      5517 non-null   object
 12  Avg Monthly GB Download            5517 non-null   float64
 13  Online Security                    5517 non-null   object
 14  Online Backup                      5517 non-null   object
 15  Device Protection Plan             5517 non-null   object
 16  Premium Tech Support               5517 non-null   object
 17  Streaming TV                       5517 non-null   object
 18  Streaming Movies                   5517 non-null   object
 19  Streaming Music                    5517 non-null   object
...
 29  Total Revenue                      7043 non-null   float64
 30  Customer Status                    7043 non-null   object
dtypes: float64(7), int64(5), object(19)
```

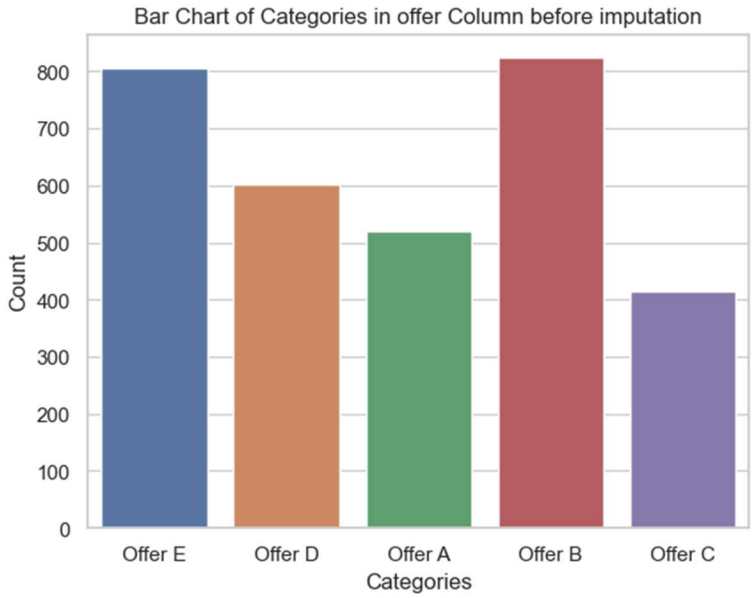
Data Visualization:



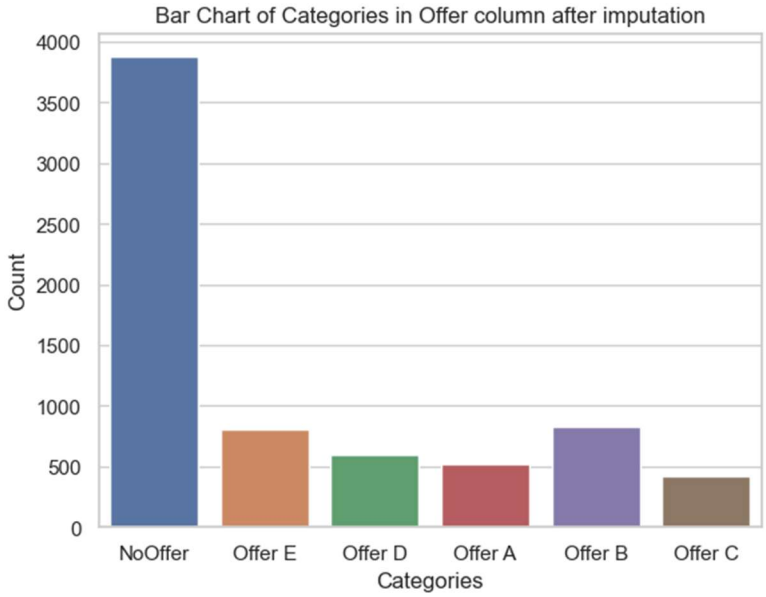
- The bar chart represents the count of different types of internet connections. It shows that Fiber Optic is the most common connection type followed by DSL and Cable, with Fiber Optic's prevalence being more than double that of Cable. Cable is the least common with less than half the count of DSL connections.



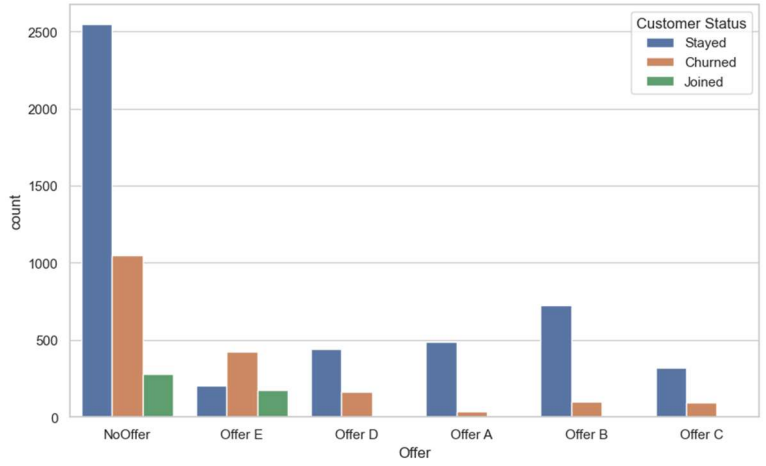
- The updated bar chart shows that after imputation, Fiber Optic remains the most common type of internet connection, Cable is still the least common, and there's a new category 'No Connection' with a count close to that of DSL.



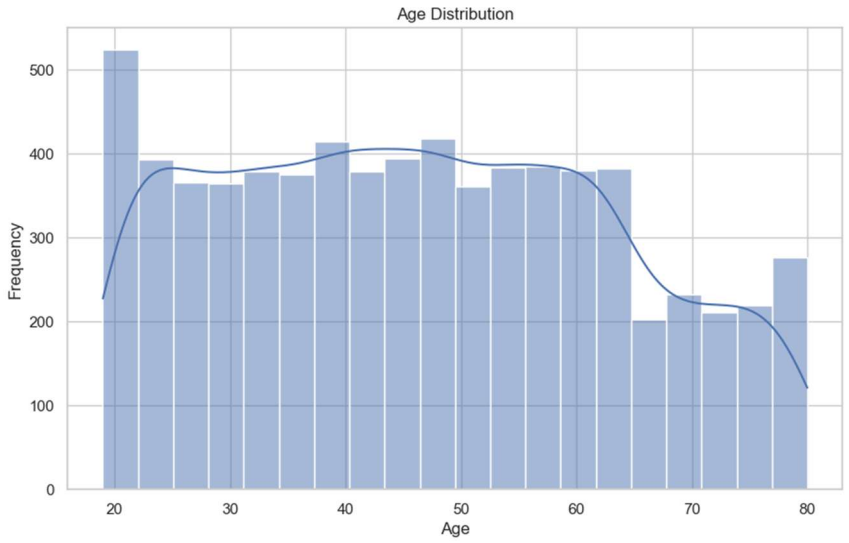
- Before imputation, Offer B is the most selected option, with Offer C being the least chosen. Offers E, D, and A have a relatively similar count, with E being slightly higher than D and A.



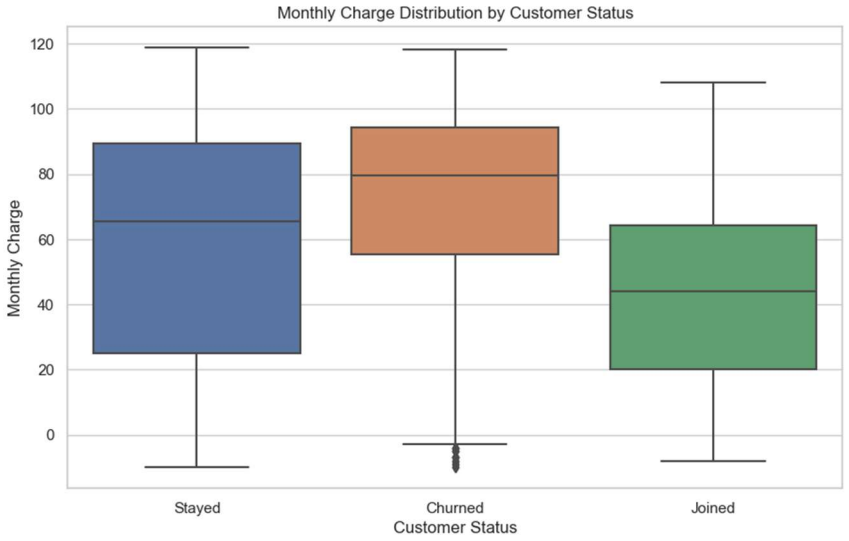
- After imputation, the category 'NoOffer' dominates, suggesting many entries had no offer selected or data was missing. The counts for Offers E, D, A, B, and C are much lower and quite similar to each other, with Offer C remaining the least frequent.



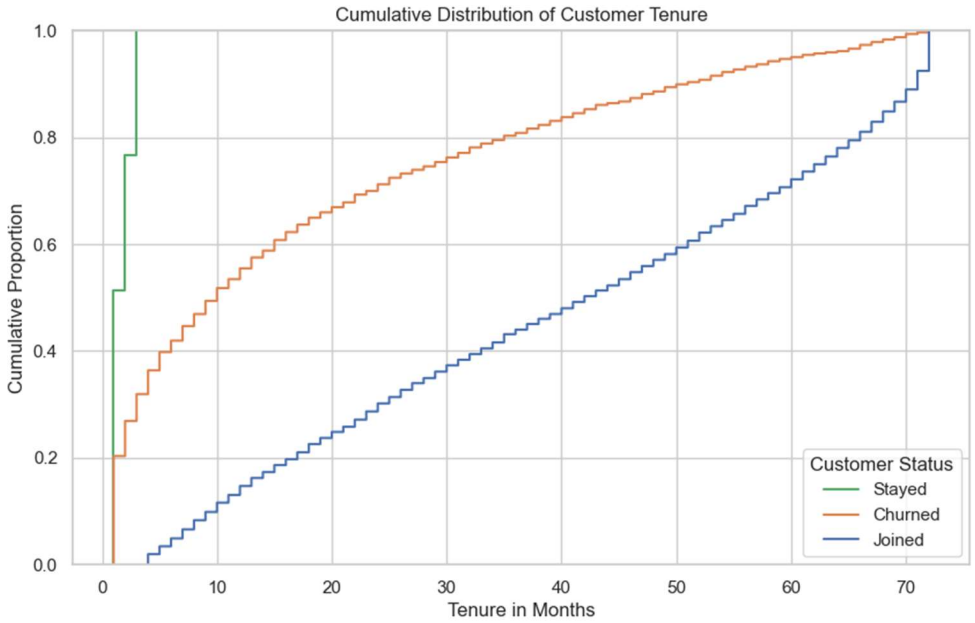
- In this bar chart, customer responses are segmented by their status—Stayed, Churned, and Joined—in relation to different offers. 'NoOffer' has the highest count, indicating most customers did not receive an offer, with a significant number of churns. Offer B appears to be the most effective in attracting new customers (Joined), while Offer C has the lowest engagement across all customer statuses. Offers E, D, and A show a mix of Stayed and Churned customers with minimal new joins.



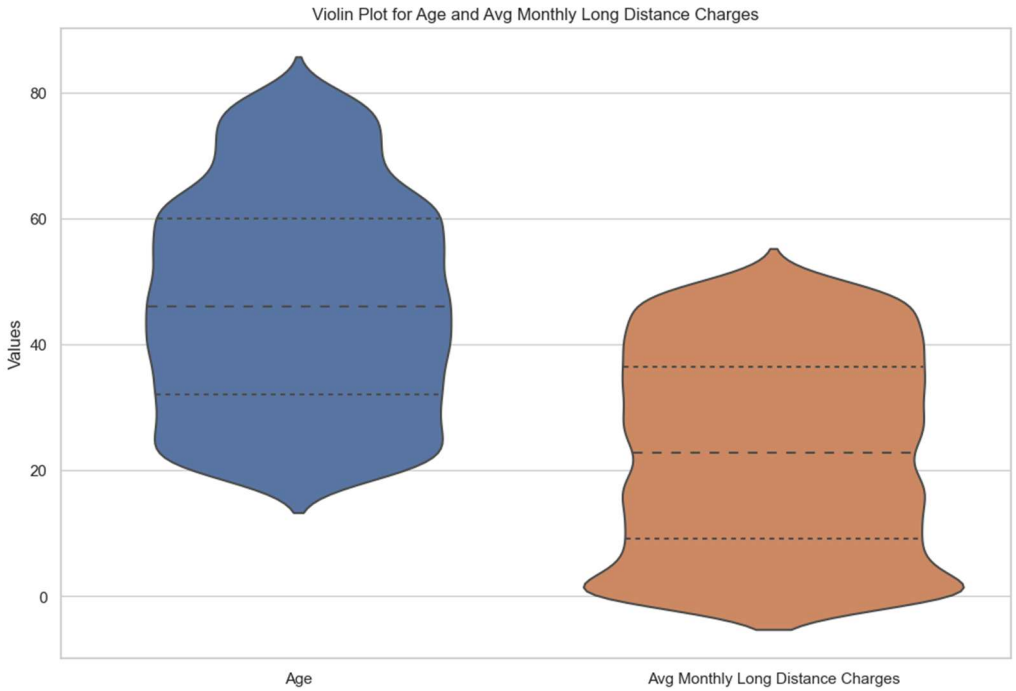
- The histogram indicates a normally distributed age demographic with a peak in early adulthood and a gradual decrease in frequency with age. There's a notable right skew showing a smaller but significant population of older adults, past the age of 70.



- The boxplot analysis of monthly charges across customer statuses reveals distinct spending patterns. Stayed customers tend to have higher and more consistent charges, while Churned customers exhibit lower and slightly more variable charges. Joined customers show diverse spending habits, possibly reflecting early-stage engagement. Understanding these patterns is crucial for segmentation, churn prediction, and targeted marketing strategies.



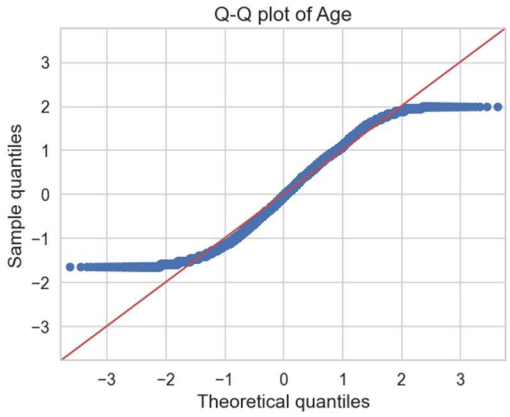
- Overall tenure: Most customers churn within the first 20 months, regardless of status. Stayed customers: Have the highest retention rate, with 50% lasting over 40 months and 20% exceeding 60 months. Churned customers: Have the lowest retention rate, with 50% churning before 10 months and almost all gone by 40 months. Joined customers: Fall somewhere in between, with 50% lasting over 20 months but not showing the same long-term retention as Stayed customers. This suggests focusing efforts on retaining Joined customers early on and understanding why Stayed customers stay loyal for longer.



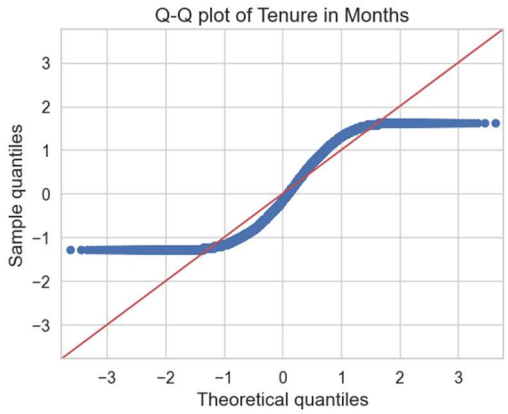
- The violin plot reveals a wider age distribution with a median around 40-50 years, skewed to the right. Avg Monthly Long Distance Charges exhibit concentrated values around 20− 30 with outliers indicating higher charges. No clear correlation between age and charges is observed, suggesting age may not strongly predict these charges.

Data Exploration:

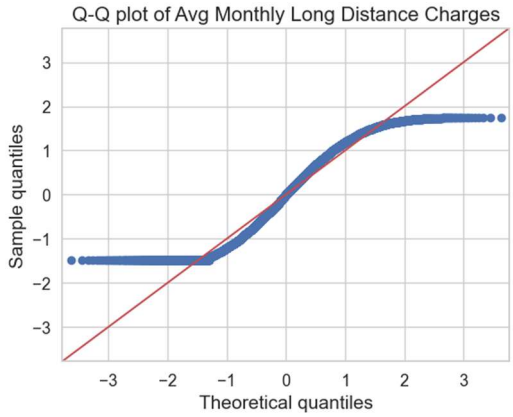
Statistical Plots:



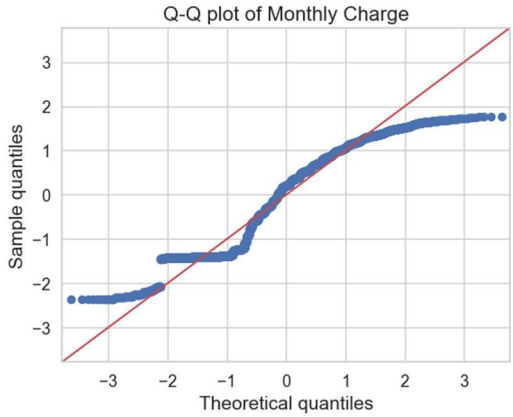
- The distribution of age closely follows a normal distribution with slight deviations at the tails.



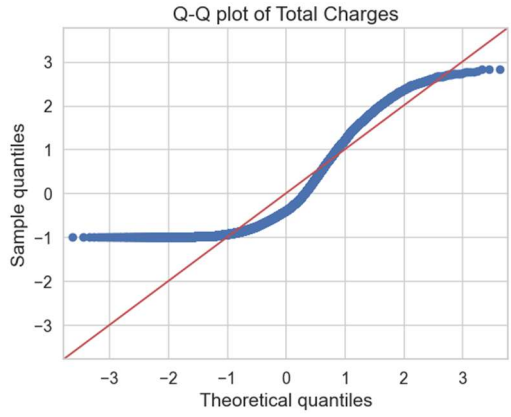
- The tenure data is normally distributed, with slight deviations at the ends.



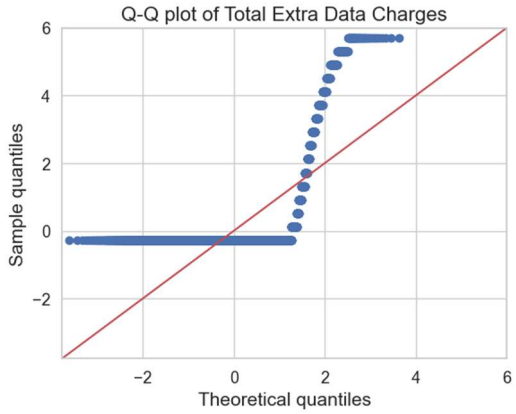
- This distribution shows a normal pattern with some deviation at the higher end.



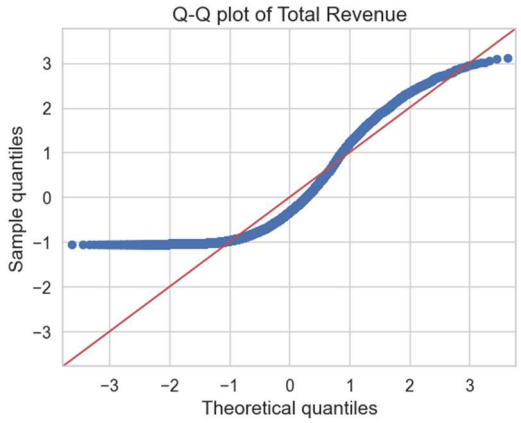
- The distribution slightly deviates from normality, especially at the lower and higher quantiles.



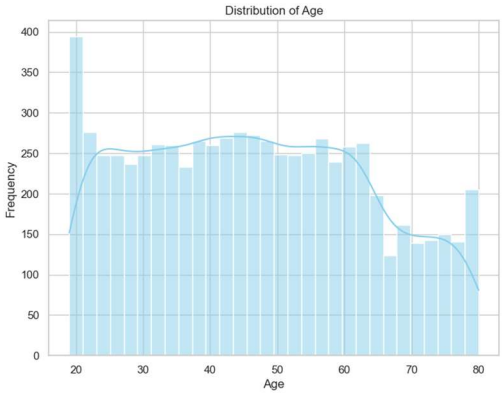
- The distribution shows a normal pattern with notable deviations at the higher end.



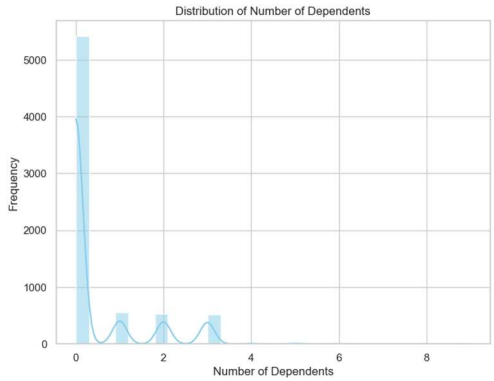
- The distribution does not follow a normal pattern, indicating a non-normal distribution of extra data charges.



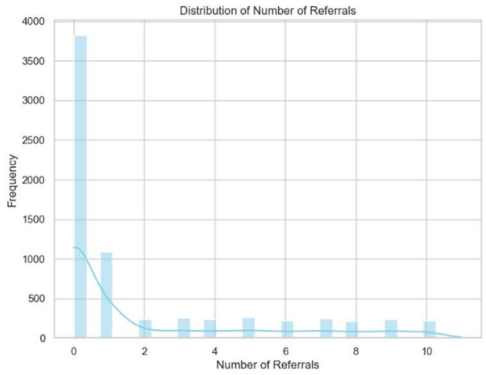
- This distribution is quite close to normal but with a slight deviation at the higher end.



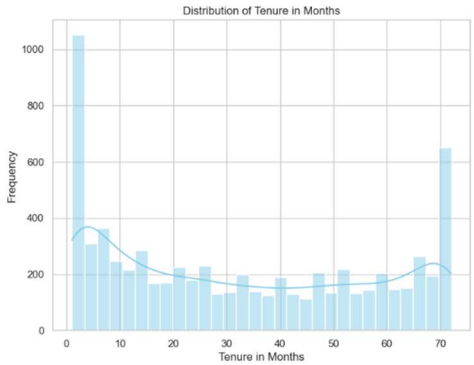
- The age distribution is skewed towards younger ages with a notable decrease after the 60s.



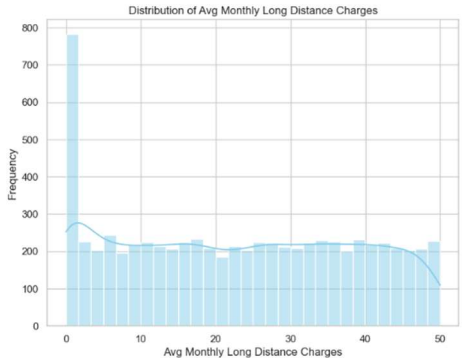
- Most individuals have zero dependents, with numbers decreasing as the number of dependents increases.



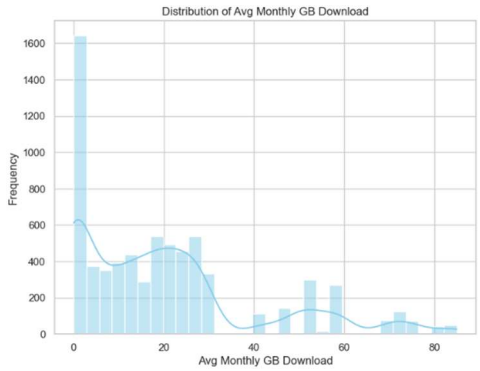
- A large number of individuals have made zero referrals, with a sharp decline in frequency for one or more referrals.



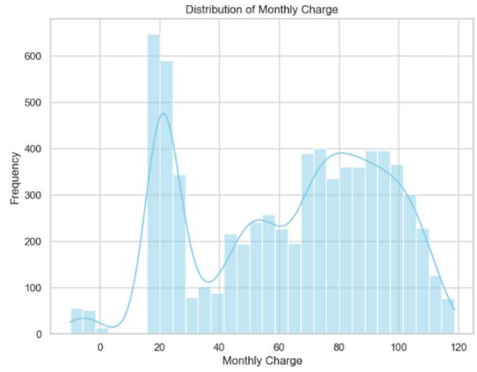
- There's a high frequency of new tenure, with a decrease as tenure length increases, and a small peak at 72 months.



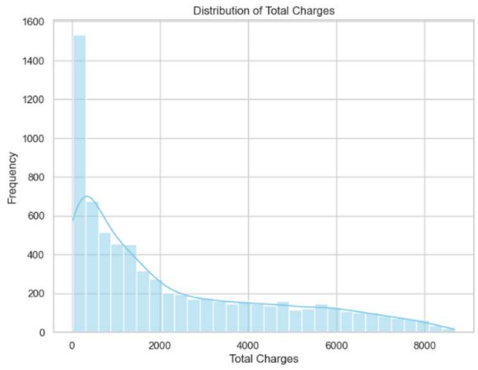
- The distribution is skewed to the right, with most people incurring low long-distance charges.



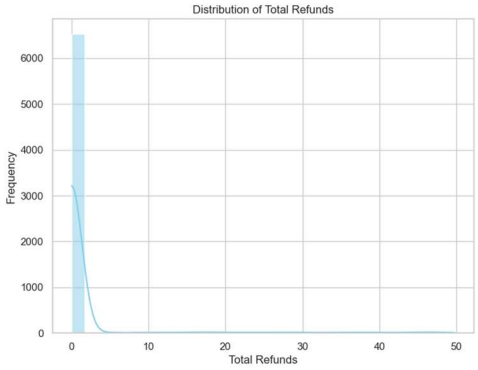
- A right-skewed distribution with a peak at low GB usage and fewer high GB users.



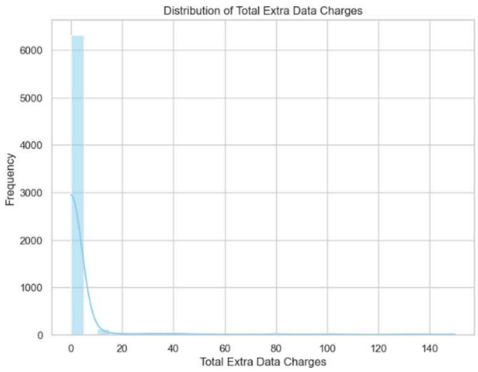
- The distribution shows multiple peaks, indicating varied pricing tiers.



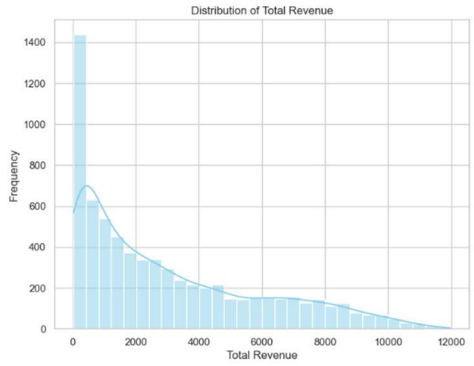
- This distribution is right-skewed, with most individuals accumulating lower total charges.



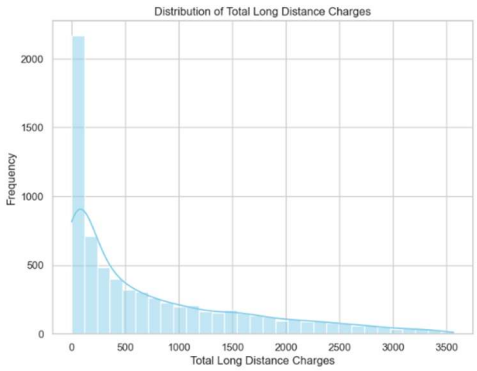
- Most individuals have not received refunds, and very few have high refund amounts.



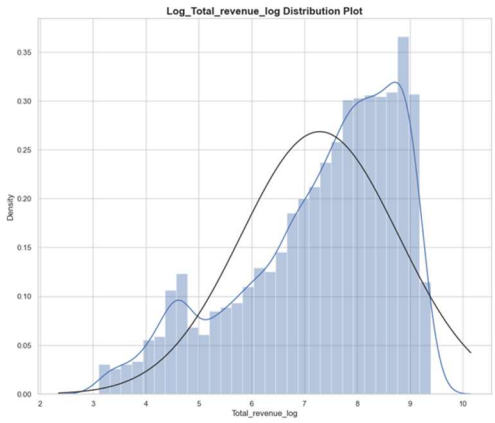
- A large number of individuals have no extra data charges, and the frequency drops as the charges increase.



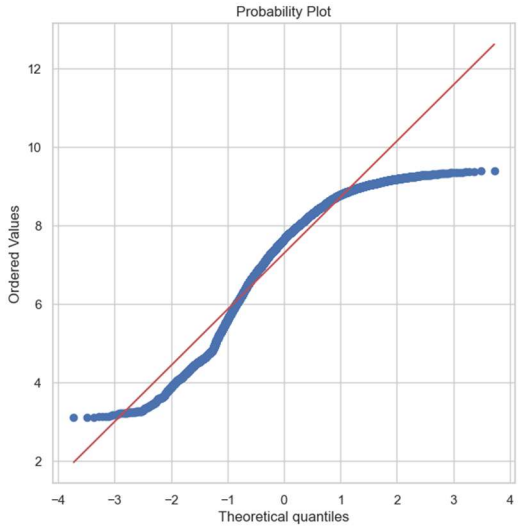
- The data is right-skewed with most individuals incurring lower long distance charges and a few incurring high charges.



- The distribution is right-skewed, indicating that most individuals generate lower total revenue, with the frequency decreasing as revenue increases.



- The distribution of the logarithm of total revenue appears to be roughly normally distributed, with the peak density around the middle of the range. There's a slight right skew, and a couple of outliers are evident at the higher end of the distribution.



- The probability plot shows that the data deviates from the theoretical line at both ends, indicating potential right-skewness or heavy tails, suggesting the data may not be normally distributed.

Encoding:

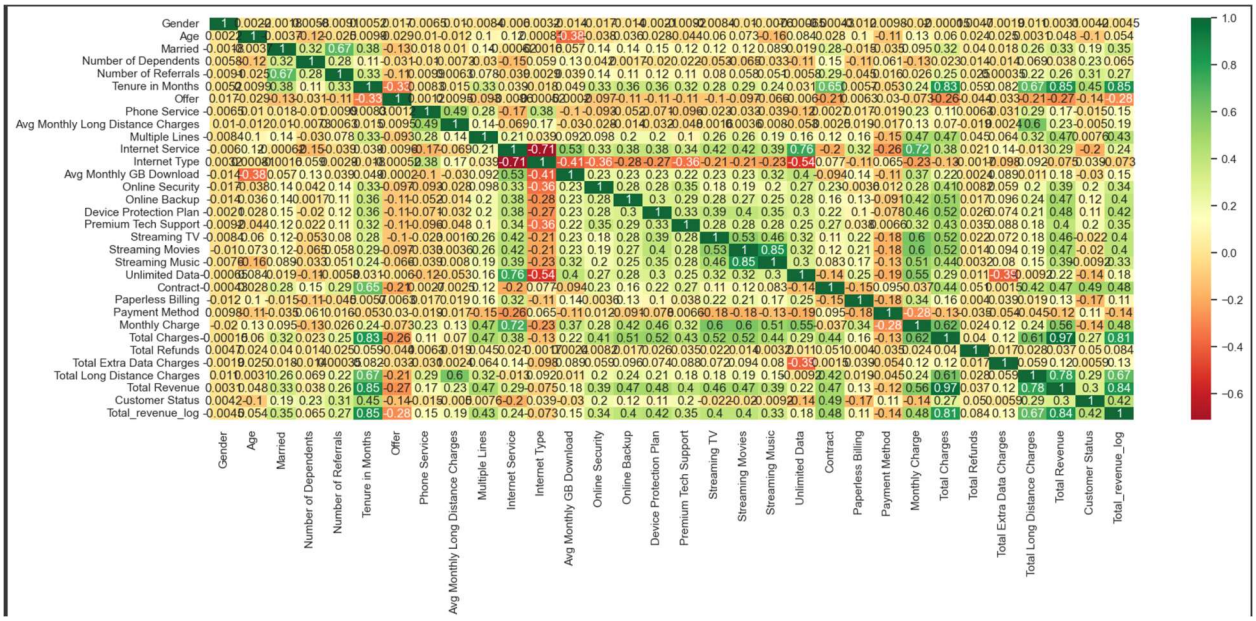
Separating all the Categorical Values and the Numerical Values to encode the categorical values for the further analysis.

	Gender	Age	Married	Number of Dependents	Number of Referrals	Tenure in Months	Offer	Phone Service	Avg Monthly Long Distance Charges	Multiple Lines	...	Paperless Billing	Payment Method	Monthly Charge	Total Charges	Total Refunds	Total Extra Data Charges	Total Long Distance Charges	Total Revenue	Customer Status	Total_revenue_log
0	0	-0.567773	1	-0.486835	0.016039	-0.952994	0	0	1.257916	0	...	1	1	0.064221	-0.744500	-0.248313	-0.273300	-0.434195	-0.718872	2	-0.275361
1	1	-0.030433	0	-0.486835	-0.650409	-0.952994	0	0	-0.794260	1	...	0	1	-2.166367	-0.766962	-0.248313	0.125055	-0.771190	-0.846108	2	-0.590140
2	1	0.208385	0	-0.486835	-0.650409	-1.156740	5	0	0.692111	0	...	1	0	0.330225	-0.882382	-0.248313	-0.273300	-0.725844	-0.914111	0	-0.848437
3	1	1.880110	1	-0.486835	-0.317185	-0.789997	4	0	0.314692	0	...	1	0	1.102599	-0.460063	-0.248313	-0.273300	-0.457641	-0.500827	0	0.057655
4	0	1.700997	1	-0.486835	0.349263	-1.197489	0	0	-1.008541	0	...	1	1	0.650712	-0.888318	-0.248313	-0.273300	-0.858681	-0.958059	0	-1.090740
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
7038	0	-1.582749	0	-0.486835	-0.650409	-0.789997	4	0	1.535639	0	...	0	1	-0.270687	-0.678482	-0.248313	-0.273300	-0.168036	-0.588006	2	-0.056536
7039	1	-0.388660	1	-0.486835	-0.317185	-0.423253	4	0	-0.437557	1	...	1	0	0.689171	-0.179466	-0.248313	-0.273300	-0.463855	-0.280726	0	0.281213
7040	1	-1.463340	0	-0.486835	-0.650409	-1.238238	5	0	-0.280893	0	...	1	1	-0.426124	-0.965390	-0.248313	-0.273300	-0.840845	-1.013748	1	-1.626880
7041	1	-1.523045	1	-0.486835	1.015710	1.410464	1	0	-1.349060	0	...	0	1	0.136331	1.035837	-0.248313	-0.273300	-0.717056	0.605693	2	0.792704
7042	1	-0.627478	1	-0.486835	-0.317185	1.247467	0	0	-1.486303	0	...	0	0	-0.147300	0.629824	-0.248313	-0.273300	-0.884833	0.234981	2	0.623213

Encoded all the categorical data thereby having a int dataset throughout. Enabling us to plot a correlation matrix.

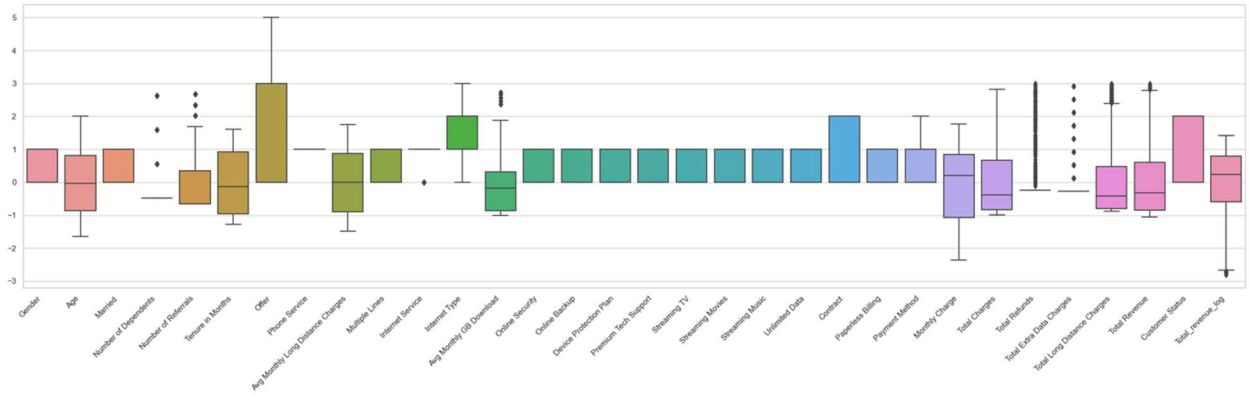
	Gender	Age	Married	Number of Dependents	Number of Referrals	Tenure in Months	Offer	Phone Service	Avg Monthly Long Distance Charges	Multiple Lines	...	Paperless Billing	Payment Method	Monthly Charge	Total Charges	Total Refunds	Total Extra Data Charges	Total Long Distance Charges
Gender	1.000000	0.002186	-0.001808	0.005840	-0.009074	0.005162	0.016956	-0.006488	0.010130	-0.008414	...	-0.011754	0.009836	-0.020073	-0.000152	0.004725	-0.001921	0.011092
Age	0.002186	1.000000	-0.003666	-0.119000	-0.025141	0.009927	-0.029311	0.009965	-0.011749	0.103122	...	0.100723	-0.112154	0.134511	0.059684	0.024168	0.025036	0.003065
Married	-0.001808	-0.003666	1.000000	0.324205	0.672867	0.380394	-0.127423	0.017706	0.010215	0.142057	...	-0.014877	-0.034503	0.094775	0.317777	0.040142	0.017882	0.264230
Number of Dependents	0.005840	-0.119000	0.324205	1.000000	0.278003	0.108237	-0.030712	-0.010149	-0.007324	-0.030307	...	-0.106814	0.061427	-0.125649	0.022535	0.014023	-0.014436	0.068966
Number of Referrals	-0.009074	-0.025141	0.672867	0.278003	1.000000	0.326975	-0.107161	0.009947	0.006336	0.078080	...	-0.044888	0.016289	0.026301	0.250378	0.024756	0.000350	0.216190
Tenure in Months	0.005162	0.009927	0.380394	0.108237	0.326975	1.000000	-0.325710	0.008271	0.014596	0.332101	...	0.005743	-0.052729	0.239065	0.826074	0.059021	0.082266	0.674149
Offer	0.016956	-0.029311	-0.127423	-0.030712	-0.107161	-0.325710	1.000000	0.001162	0.009475	-0.093313	...	0.006323	0.030120	-0.072815	-0.263946	-0.043731	-0.032732	-0.212498
Phone Service	-0.006488	0.009965	0.017706	-0.010149	0.009947	0.008271	0.001162	1.000000	0.486673	0.279690	...	0.016505	-0.018615	0.234170	0.113106	0.006331	-0.030620	0.289728
Avg Monthly Long Distance Charges	0.010130	-0.011749	0.010215	-0.007324	0.006336	0.014596	0.009475	0.486673	1.000000	0.136004	...	0.018673	-0.016536	0.130087	0.069500	-0.018644	0.002414	0.599828
Multiple Lines	-0.008414	0.103122	0.142057	-0.030307	0.078080	0.332101	-0.093313	0.279690	0.136004	1.000000	...	0.163530	-0.152733	0.467398	0.468615	0.045491	0.064443	0.323165
Internet Service	-0.006026	0.117346	-0.000615	-0.153137	-0.038984	0.038582	-0.009553	-0.172209	-0.068605	0.210564	...	0.321013	-0.260186	0.724477	0.375269	0.020515	0.143736	-0.013128
Internet Type	0.003217	0.000809	-0.001627	0.059232	0.002913	-0.017825	-0.000520	0.382482	0.171752	0.038916	...	-0.110804	0.065033	-0.232379	-0.128057	-0.001744	-0.097967	0.091938
Avg Monthly GB Download	-0.014065	-0.376595	0.056745	0.129966	0.038575	0.049119	-0.000197	-0.102748	-0.030455	0.091679	...	0.142999	-0.111482	0.372776	0.223268	0.002397	0.088939	0.011373
Online Security	-0.017021	-0.038001	0.143106	0.041977	0.142416	0.327543	-0.096948	-0.092893	-0.028108	0.098108	...	-0.003636	0.011637	0.282875	0.411922	0.008190	0.058889	0.200477
Online Backup	-0.013773	0.035541	0.141498	0.001721	0.112369	0.360558	-0.108924	-0.052312	-0.013909	0.202237	...	0.126735	-0.090512	0.420707	0.509507	0.016855	0.095852	0.239873
Device Protection Plan	-0.002105	0.028491	0.153786	-0.019925	0.116695	0.360935	-0.112303	-0.071227	-0.031869	0.201137	...	0.103797	-0.078359	0.460685	0.522255	0.026038	0.073985	0.210057
Premium Tech Support	-0.009212	-0.043760	0.119999	0.022094	0.107725	0.324554	-0.113353	-0.096340	-0.048217	0.100571	...	0.037880	0.006641	0.318141	0.432146	0.034823	0.087941	0.182076
Streaming TV	-0.008393	0.059760	0.124666	-0.052981	0.079687	0.279928	-0.103505	-0.022574	0.001641	0.257152	...	0.223841	-0.180426	0.599838	0.515217	0.021796	0.072300	0.182302

Heatmap:



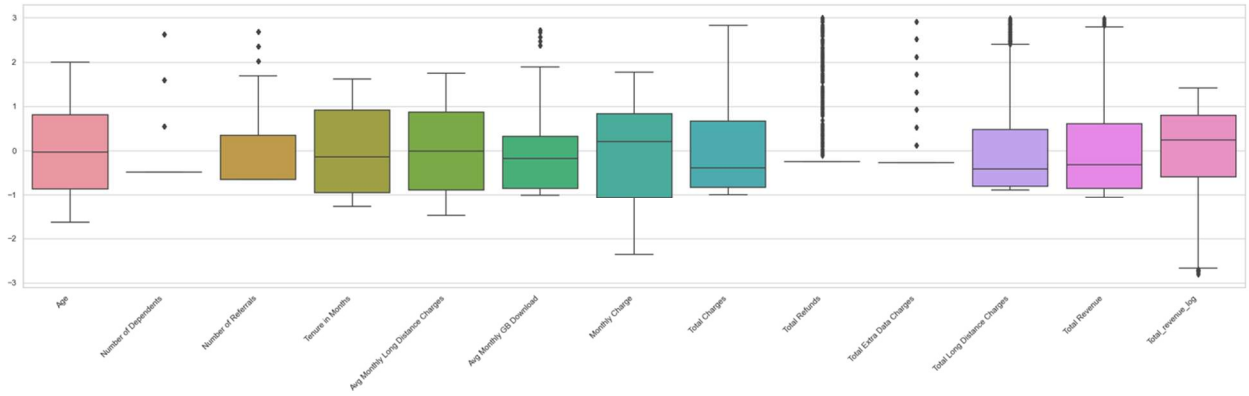
- The image displays a heatmap of correlation coefficients between various variables. Dark red signifies a strong negative correlation, dark green indicates a strong positive correlation, and colors closer to beige represent little to no correlation. For example, 'Total Revenue' and 'Total Charges' have a high positive correlation, indicated by the dark green square. In contrast, 'Contract' and 'Customer Status' have a strong negative correlation, as shown by the dark red square.

Normalization:

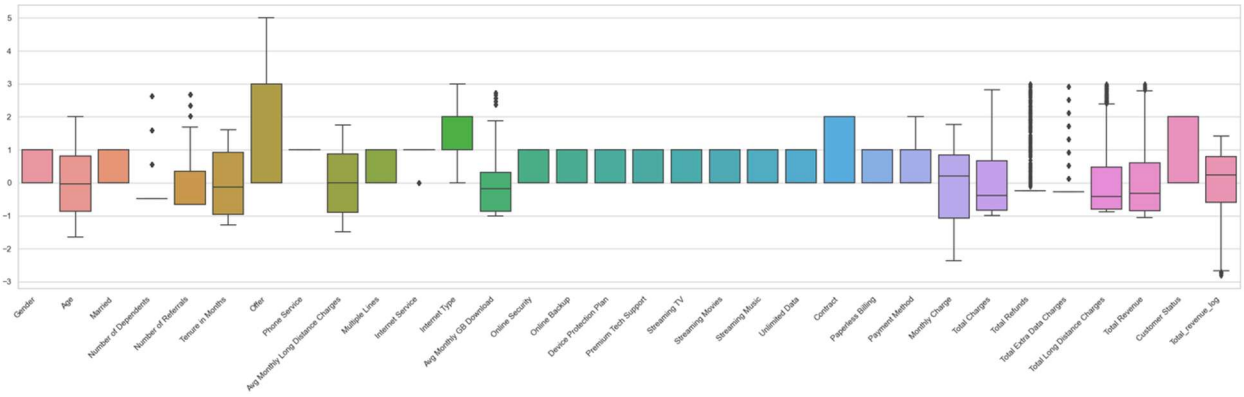


- This boxplot visualizes the distribution of various transformed variables, likely standardized as they are centered around zero. The spread of the data for each variable is shown by the length of the boxes and whiskers. For example, 'Total Refunds' has a wide range of values with many outliers, while 'Married' has a more compact distribution with no outliers. Variables like 'Total Revenue' and 'Customer Status' also have wider spreads, indicating greater variability in their data compared to others.

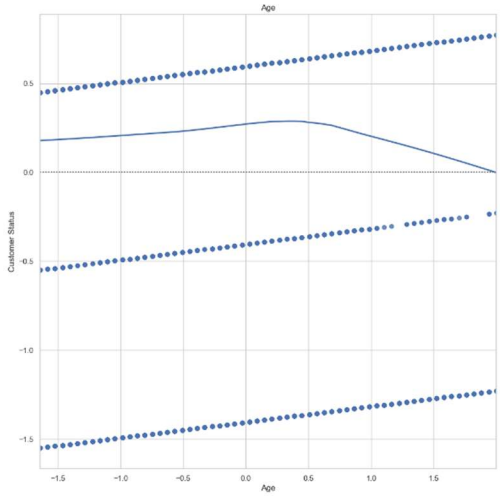
Handling Outliers:



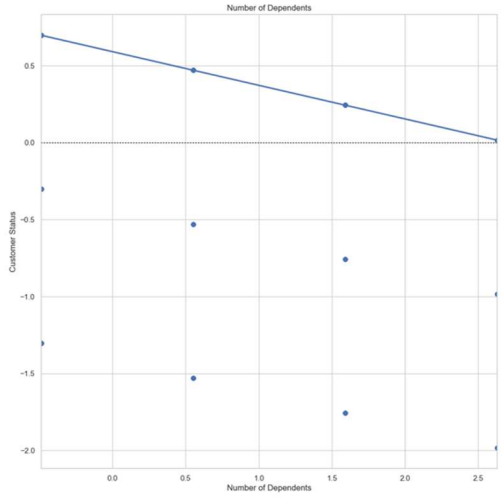
- This boxplot graphically depicts the distributions of several variables, which seem to be normalized or standardized (as values are centered around zero). The variables display varying degrees of spread and outliers. For instance, 'Total Refunds' has a narrow interquartile range but many outliers, whereas 'Age' and 'Number of Dependents' show a broader spread with fewer outliers. The tails of the whiskers indicate the range of the data, and points outside the whiskers are plotted as individual points representing outliers.



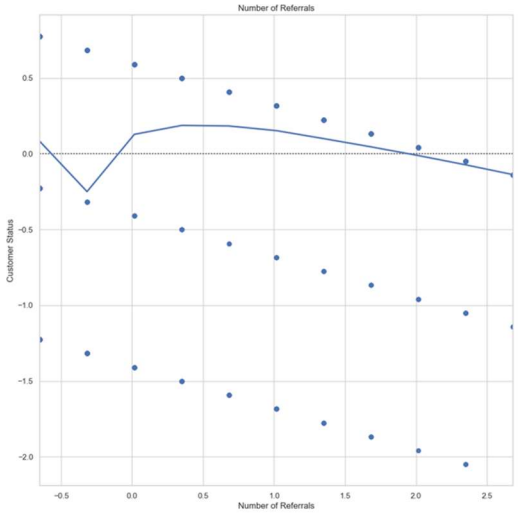
- The boxplot illustrates the distributions of standardized or normalized variables, where the central line in each box represents the median. The box length indicates the interquartile range (IQR), and the whiskers extend to show the rest of the distribution, except for outliers, which are plotted as individual points. Notably, 'Phone Service' and 'Contract' have wider IQRs indicating more variability, while 'Offer' has a smaller IQR, suggesting less variability. Outliers are present in several variables, most prominently in 'Total Refunds'.



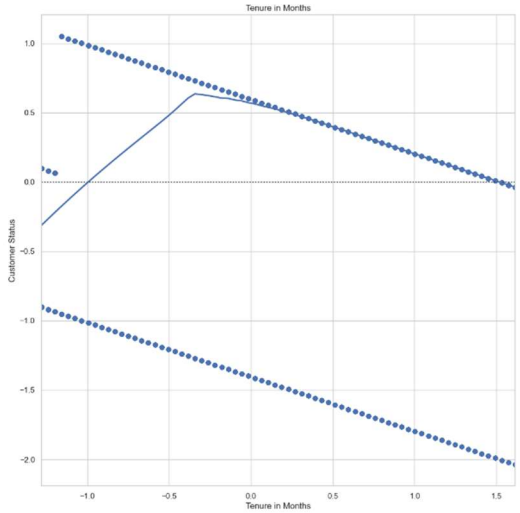
- The scatter plot with a fitted line suggests a relationship between age (possibly standardized) and customer status. There is no strong linear correlation as the points are widely scattered and the fitted line is relatively flat. The distribution of points does not indicate any clear pattern or trend between age and customer status.



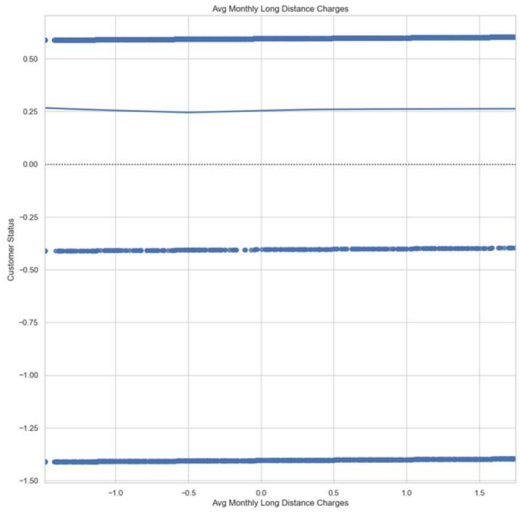
- The scatter plot with a fitted line indicates a potential negative relationship between the number of dependents and customer status, with the line sloping downwards as the number of dependents increases. However, the data points are sparse and widely spread out, implying that this trend should be interpreted with caution due to the limited data and potential presence of outliers.



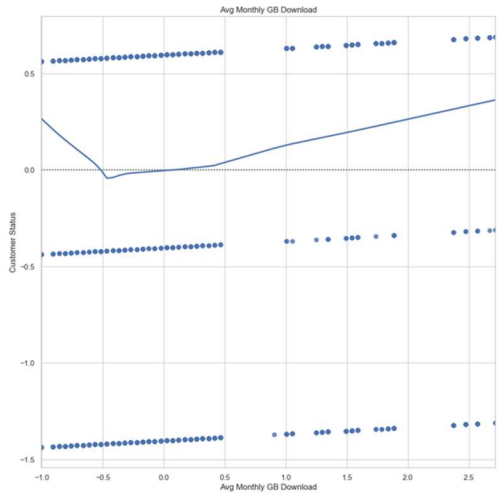
- The scatter plot and fitted line suggest that the number of referrals has a non-linear relationship with customer status, as indicated by the curve in the fitted line. The plot shows some variability in customer status across different numbers of referrals, but there is not a clear or strong linear trend. There are a few outliers, especially for customers with a higher number of referrals.



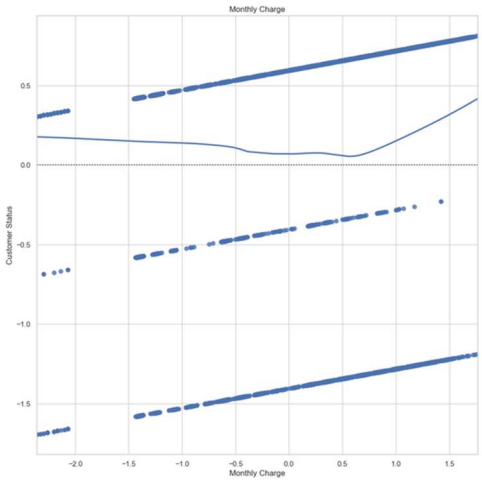
- The scatter plot with a fitted line suggests a negative relationship between tenure in months and customer status, with the line sloping downwards as tenure increases. This could indicate that longer tenure might be associated with a lower (possibly negative) customer status, but this observation would require further statistical analysis to confirm.



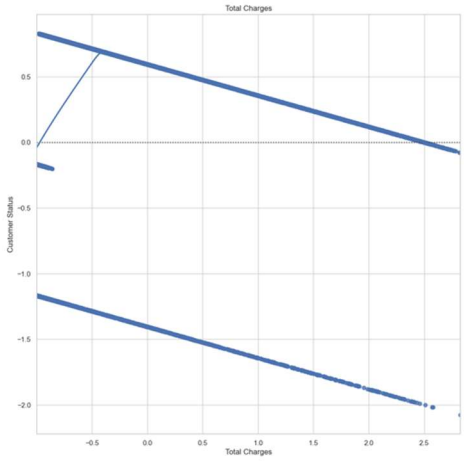
- The scatter plot shows no clear trend or correlation between average monthly long-distance charges and customer status, as indicated by the flat line and the spread of points along the horizontal axis.



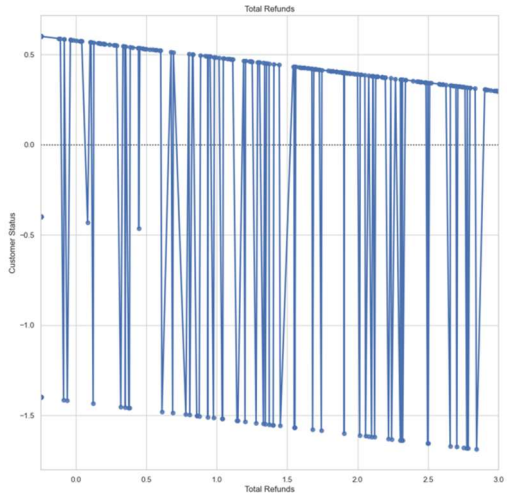
- The scatter plot indicates a non-linear relationship between average monthly GB download and customer status, with no apparent trend across the range of data.



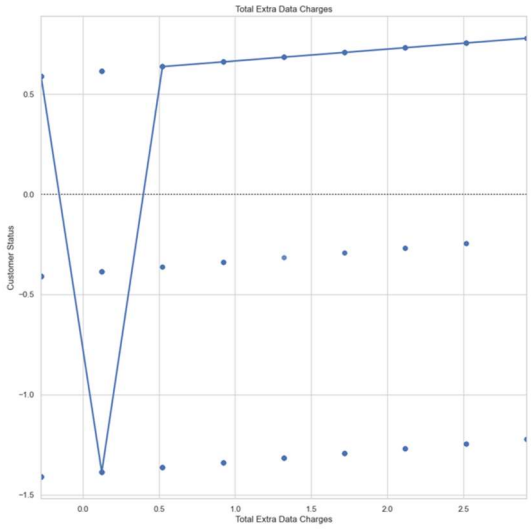
- The scatter plot and fitted line suggest a positive relationship between monthly charge and customer status, with customer status increasing as monthly charges increase. The pattern is consistent across the range of monthly charge values.



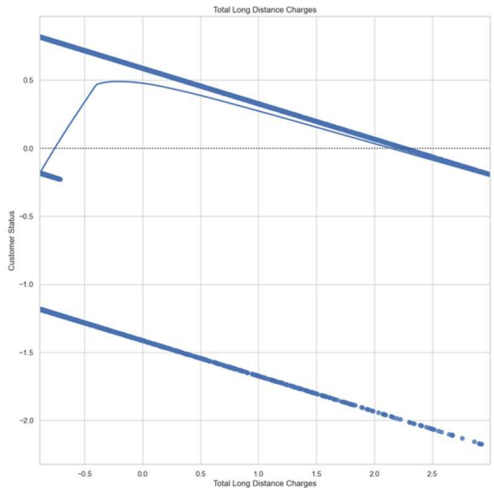
- The scatter plot indicates a negative relationship between total charges and customer status, with customer status decreasing as total charges increase. The trend is downward-sloping across the range of total charge values.



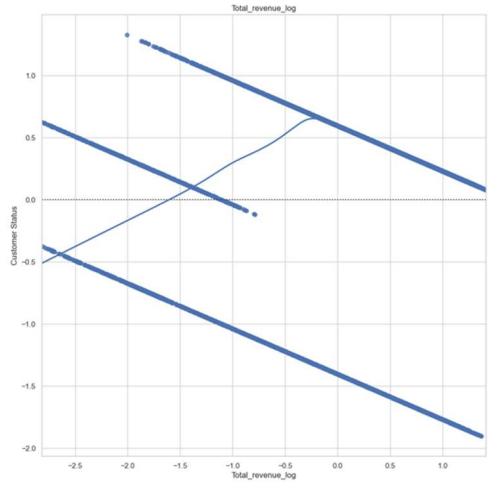
- The scatter plot shows distinct vertical lines at specific points along the 'Total Refunds' axis, suggesting that the refund amounts are discrete and not continuous. There is no clear trend or relationship visible between total refunds and customer status; customer status varies across the range of total refunds without a discernible pattern.



- The scatter plot shows a V-shaped relationship between total extra data charges and customer status, indicating that customer status is higher with no or high extra data charges and lower with moderate extra data charges.

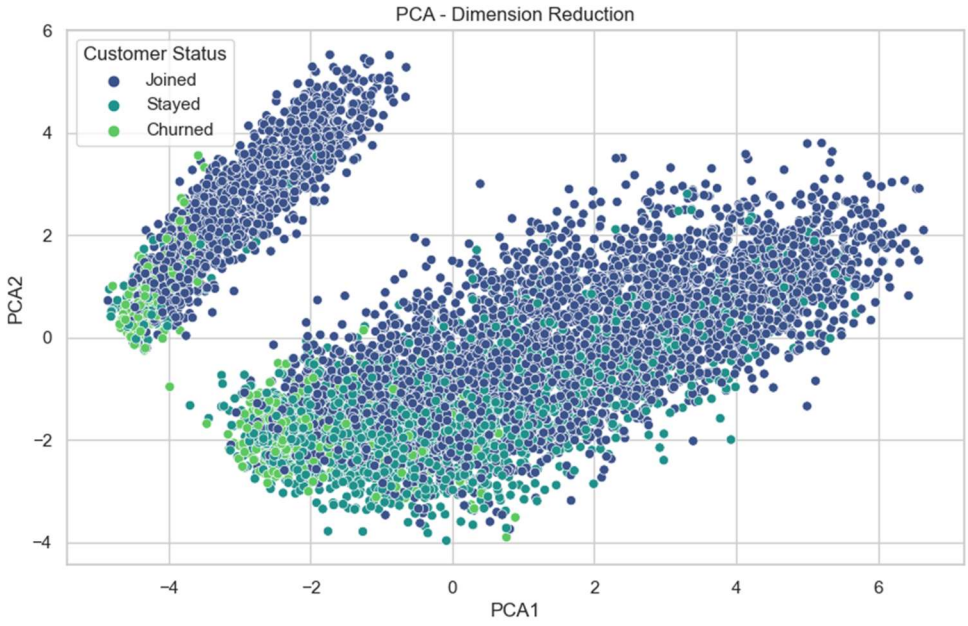


- The plot indicates a generally negative relationship between total long distance charges and customer status, with status decreasing as charges increase.



- The scatter plot shows a negative relationship between the logarithm of total revenue and customer status, with status decreasing as log revenue increases.

Dimension Reduction:



- The PCA (Principal Component Analysis) scatter plot illustrates the distribution of customer status—'Joined', 'Stayed', 'Churned'—along the first two principal components. The clusters for each status overlap significantly, suggesting that PCA has extracted components that do not linearly separate the different statuses well. The plot also indicates that there is variability among the customer statuses that is not captured by the first two principal components alone.

Model Exploration and Finalization

In the exploration of candidate data mining models for customer churn prediction in a telecommunications company's dataset (Q2 2022), several models have been considered. Here are brief descriptions for each model:

1. Logistic Regression:

- Logistic Regression is chosen for its simplicity and interpretability. It works well when the relationship between features and churn is approximately linear. Logistic Regression can serve as a baseline model, providing insights into initial patterns in customer behavior.

2. Random Forest:

- Random Forest, an ensemble method, is selected due to its ability to handle non-linear relationships and capture complex patterns. With a diverse set of features such as demographics and subscription services, Random Forest is expected to identify significant predictors for customer churn effectively.

3. Gradient Boosting (XGBoost or LightGBM):

- Gradient Boosting models, specifically XGBoost or LightGBM, are considered for their capacity to capture intricate relationships and sequential learning. These models can handle both linear and non-linear patterns, making them well-suited for customer churn prediction in a dataset with diverse features.

4. Support Vector Machines (SVM):

- SVM is effective when dealing with complex decision boundaries. It works well for datasets with clear margins of separation between classes. In the context of customer churn, SVM could be a good choice if there are distinct patterns that can be linearly or non-linearly separated.

5. k-Nearest Neighbors (k-NN):

- k-NN classifies data points based on the majority class among their k-nearest neighbors. It's effective for datasets with localized patterns. In customer churn prediction, k-NN could work well if customers with similar characteristics tend to exhibit similar churn behavior.

6. Decision Tree:

- Decision Trees are intuitive and can capture non-linear relationships in the data. They are suitable for datasets with complex decision boundaries. In the context of customer churn, Decision Trees can help identify key factors influencing churn decisions and provide interpretable insights.

7. Gaussian Naive Bayes (NB):

- Gaussian NB assumes that features are conditionally independent given the class label. It's suitable for datasets with continuous features. If features in the customer churn dataset are approximately independent given churn status, Gaussian NB might provide efficient predictions.

For a brief overview of the dataset characteristics corresponding to each model:

- Logistic Regression: Suitable for datasets with approximately linear relationships between features and churn. Efficient for initial exploration and interpretation.

- Random Forest: Effective for datasets with non-linear relationships and diverse feature sets. Robust in capturing complex patterns in customer behavior.

- Gradient Boosting (XGBoost or LightGBM): Ideal for datasets with intricate relationships and sequential dependencies. Capable of handling both linear and non-linear patterns.

- Support Vector Machines (SVM): Effective for datasets with clear margins of separation between classes. Suitable for scenarios where distinct patterns need to be separated.

- k-Nearest Neighbors (k-NN): Suitable for datasets with localized patterns. Effective when customers with similar characteristics tend to have similar churn behaviors.

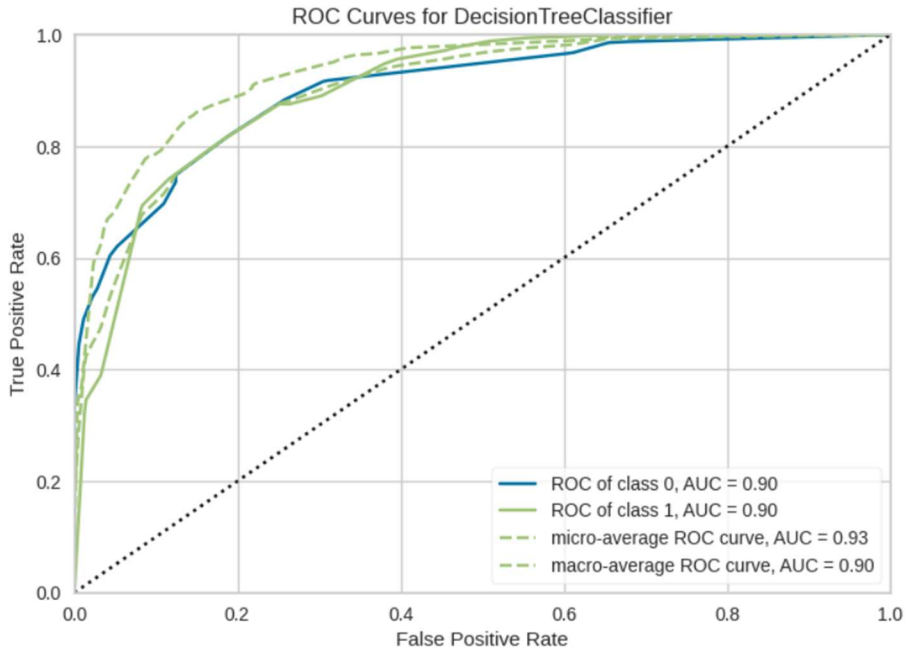
- Decision Tree: Intuitive and suitable for datasets with complex decision boundaries. Provides insights into key factors influencing churn decisions.

- Gaussian Naive Bayes (NB): Suitable for datasets with approximately independent features. Provides quick and interpretable predictions for customer churn.

Consideration factors include model performance metrics, interpretability, and computational efficiency. Ensemble methods, such as Random Forest and Gradient Boosting, may offer high predictive accuracy, while Logistic Regression and Decision Trees provide interpretability.

Model Performance Evaluation and Interpretation

Decision Tree Classifier:



Training accuracy: 100%

Testing accuracy: 83%

Results on test data:

Test Accuracy 85.48%

Test Precision 86.30%

Test Recall 94.77%

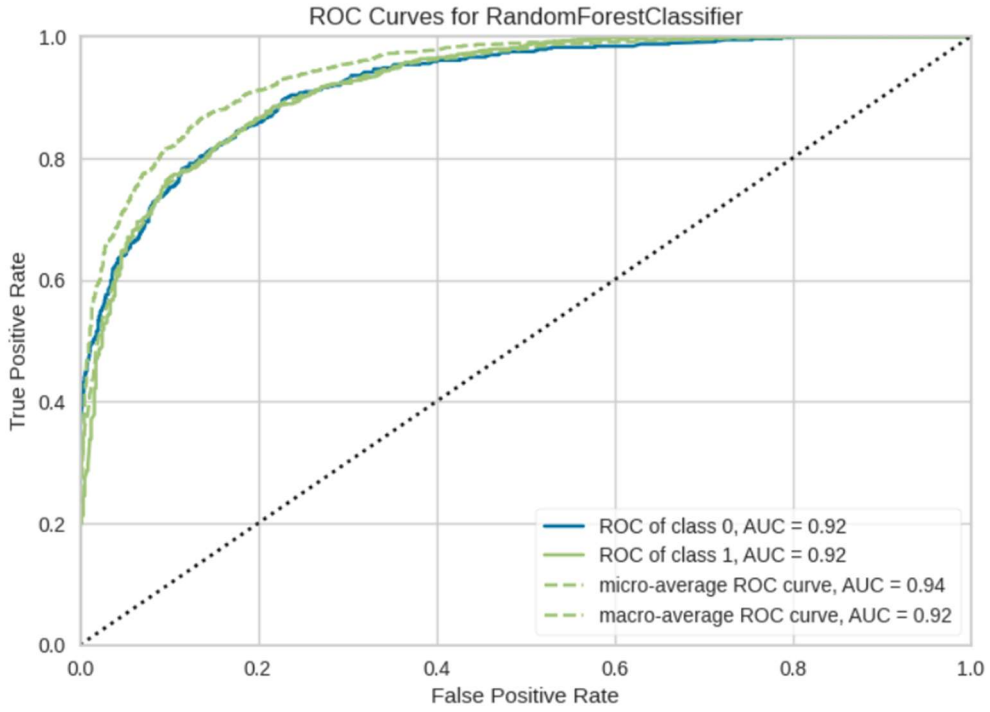
	Predicted C1	Predicted C2
Actual C1	348	213
Actual C1	74	1342

The performance of the decision tree classifier was evaluated using Receiver Operating Characteristic (ROC) curves. The ROC curves demonstrate that the decision tree classifier

performs well. While not perfectly aligned with the ideal path, the curves are relatively close, indicating good classification performance.

The presented ROC curves depict the performance for two separate classes, along with macro and micro-averaged curves. Macro-averaging treats all classes equally, while micro-averaging weights each class based on its prevalence in the data. In this case, all AUCs hover around 0.9, suggesting strong classification performance by the decision tree model for both classes.

Random Forest Model:



	Predicted C1	Predicted C2
Actual C1	359	202
Actual C2	65	1351

Training accuracy: 100

Testing accuracy: 87.2%

Results on test data:

Test accuracy: 86.49%

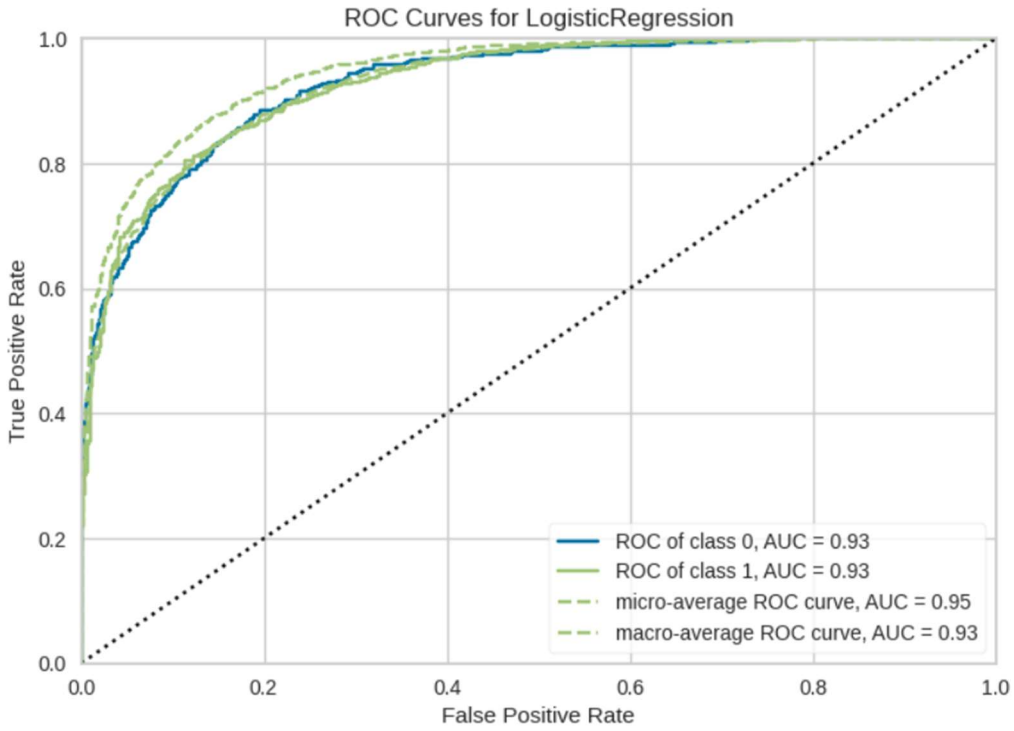
Test precision: 86.99%

Test recall: 95.41%

The ROC curves themselves don't follow the ideal path, but they are relatively close. This means the decision tree classifier is performing well. The area under the curve (AUC) is a numerical metric that summarizes the performance of a classification model.

The ROC curves are for two different classes, along with macro and micro-averaged curves. Macro-averaging treats all classes equally, while micro-averaging weighs the contribution of each class according to its presence in the data. In this case, all the AUCs are around 0.9, which suggests that the decision tree classifier is performing well at classifying instances for both classes.

Logistic Regression:



Training accuracy: 0.866

Testing accuracy: 0.863

Results on test data:

Test accuracy: 86.44%

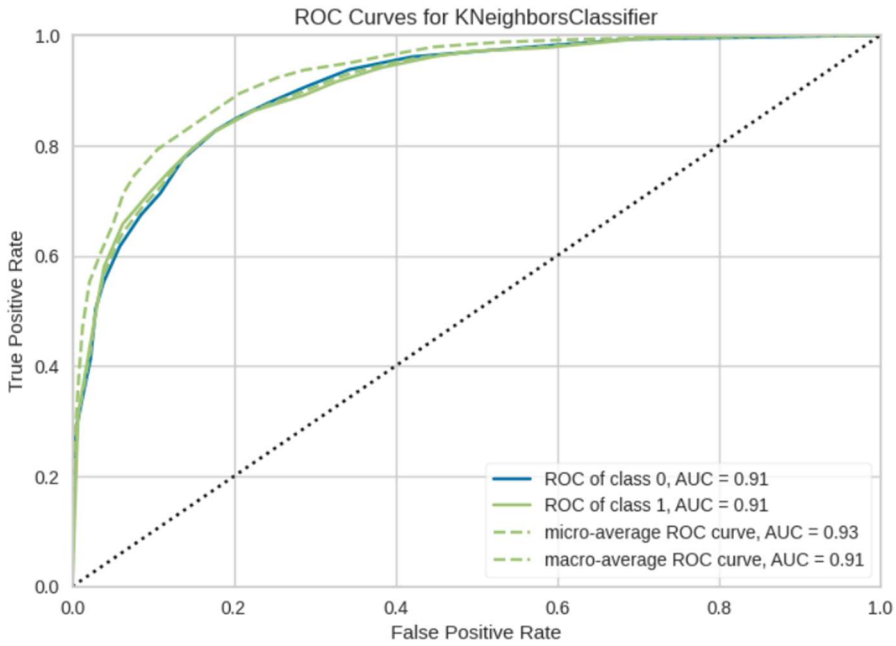
Test precision: 89.48%

Test recall: 91.88%

	Predicted C1	Predicted C2
Actual C1	408	153
Actual C2	115	1301

From the ROC curves we can see that the AUC (Area Under the Curve) for all the curves is around 0.93, which suggests that the logistic regression model is performing well at classifying instances for both classes. An AUC of 1 corresponds to a perfect classifier. The Macro-averaging treats all classes equally, while micro-averaging weights the contribution of each class according to its presence in the data.

KNN Model:



Training accuracy: 88.7%

Testing accuracy: 84.4%

Results on test data:

Test accuracy: 84.07%

Test precision: 88.69%

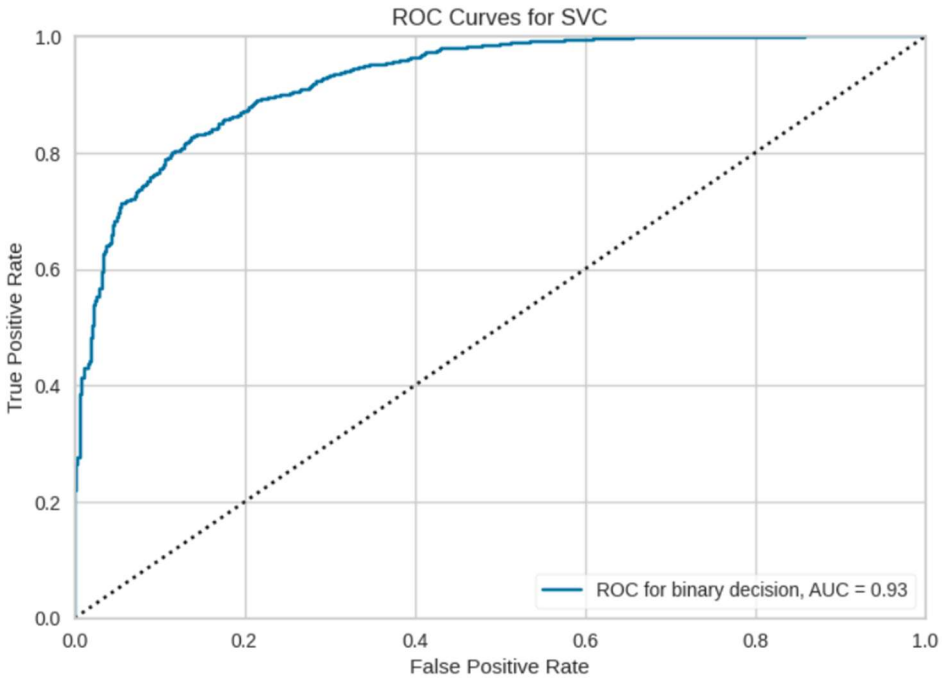
Test recall: 89.12%

	Predicted C1	Predicted C2
Actual C1	400	161
Actual C2	154	1262

The ROC curves depict the performance for four different classes. The AUC (Area Under the Curve) for each curve is around 0.91, which suggests that the KNN classifier is performing well at classifying instances for all four classes. An AUC of 1 corresponds to a perfect classifier.

In this case, both the macro and micro-average AUCs are around 0.91, signifying good overall performance by the KNN model.

Support Vector Machines:



Training accuracy: 88.3%

Testing accuracy: 86.6%

Results on test data:

Test accuracy: 86.29%

Test precision: 88.92%

Test recall: 92.37%

	Predicted C1	Predicted C2
Actual C1	398	163
Actual C2	108	1308

The ROC curve shows the performance of the SVC model on a binary decision task. The AUC (Area Under the Curve) value is 0.93, which suggests that the model is performing well. An AUC of 1 corresponds to a perfect classifier.

Performance Comparison:

	Algorithm	ROC AUC	Accuracy	Precision	f1 Score
4	Random Forest	86.84	92.09	86.80	86.97
1	Kernel SVM	86.19	91.30	86.19	86.19
0	Logistic Regression	86.01	92.24	86.01	86.01
2	KNN	82.61	86.73	82.61	82.61
5	Decision Tree Classifier	81.81	78.22	81.53	82.11
3	Gaussian NB	80.23	88.12	80.23	80.23

Based on the metrics in the table, Logistic Regression appears to be the best performing model overall. It achieves a good balance between precision, recall, and overall accuracy, with the highest ROC AUC suggesting strong ability to distinguish between positive and negative cases. Decision Tree seems to be the least favorable model due to its lower performance across all metrics.

It's important to note that the best model choice for a project depends on the specific requirements. If precision is critical, KNN or Decision Tree might be good options due to their acceptable precision values. However, if overall accuracy and balanced performance are most important, Logistic Regression appears to be the best option based on this dataset.

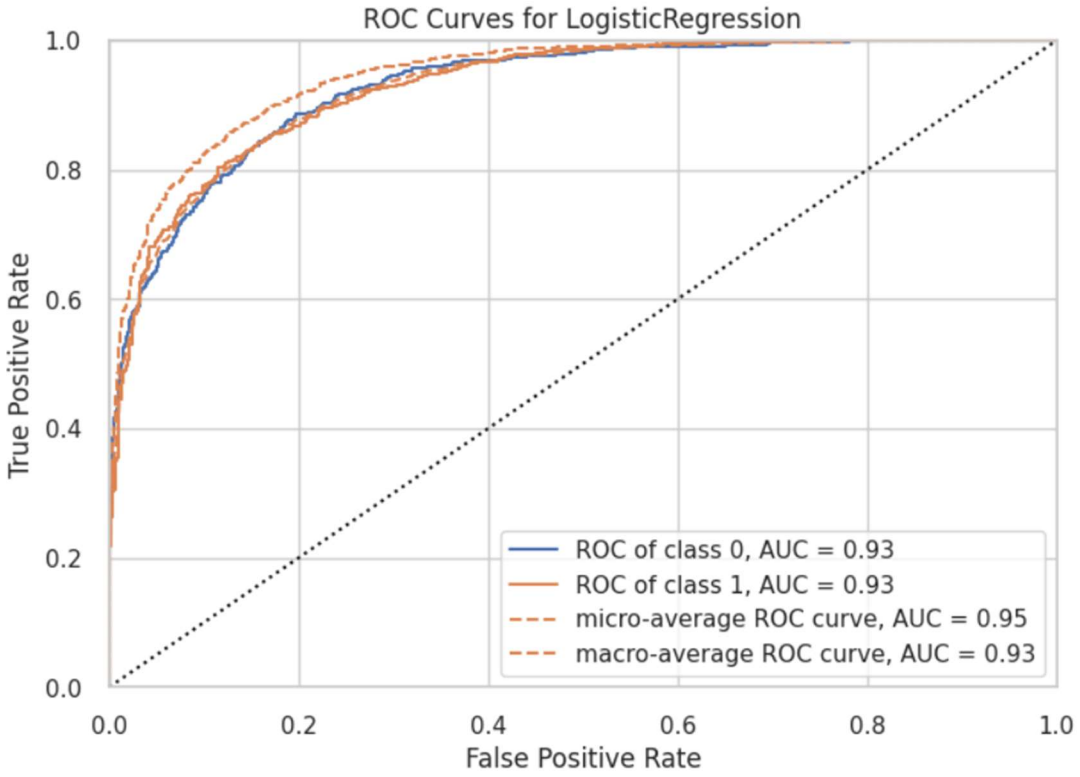
Model Training without Optimization:

Algorithm	ROC AUC	Accuracy	Precision	F1 Score
Random Forest	86.84	91.99	86.75	86.88
Kernel SVM	86.19	91.30	86.19	86.19
Logistic Regression	86.01	92.24	86.01	86.01
KNN	82.61	86.73	82.61	82.61
Decision Tree Classifier	81.85	78.07	81.77	81.94

The comparison of performance evaluation of various machine learning models before and after optimization using accuracy, precision, F1 score, and ROC AUC metrics as expected, exhibited lower performance metrics before optimization. This demonstrates that the optimization techniques were successful in improving the performance of all models.

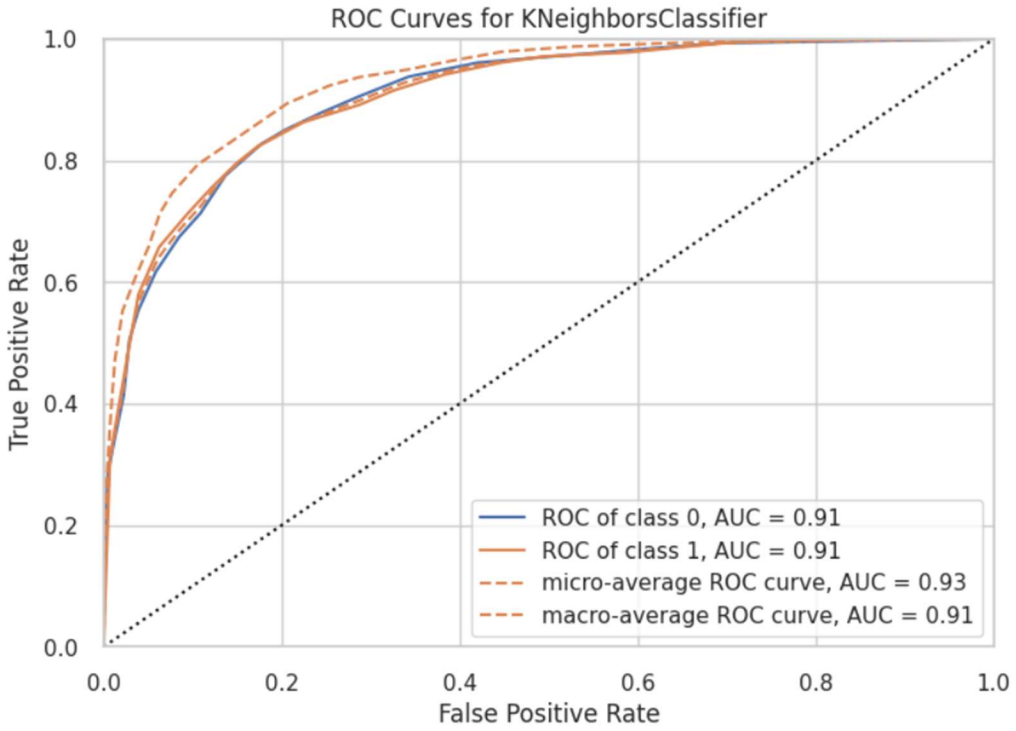
Tuning the parameters:

Logistic Regression:



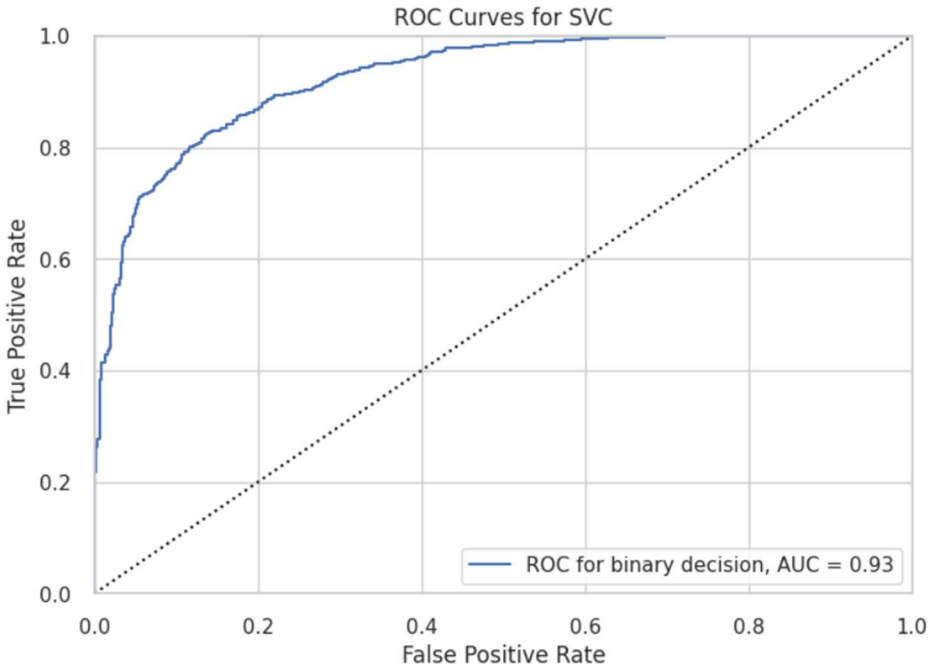
The Receiver Operating Characteristic (ROC) curves indicate that the Logistic Regression classifier has a strong ability to differentiate between two classes (0 and 1), with both showing an Area Under the Curve (AUC) of 0.93. The micro- and macro-average ROC curves suggest consistent model performance across classes, with particularly high accuracy as denoted by the micro-average AUC of 0.95.

KNN Classifier:



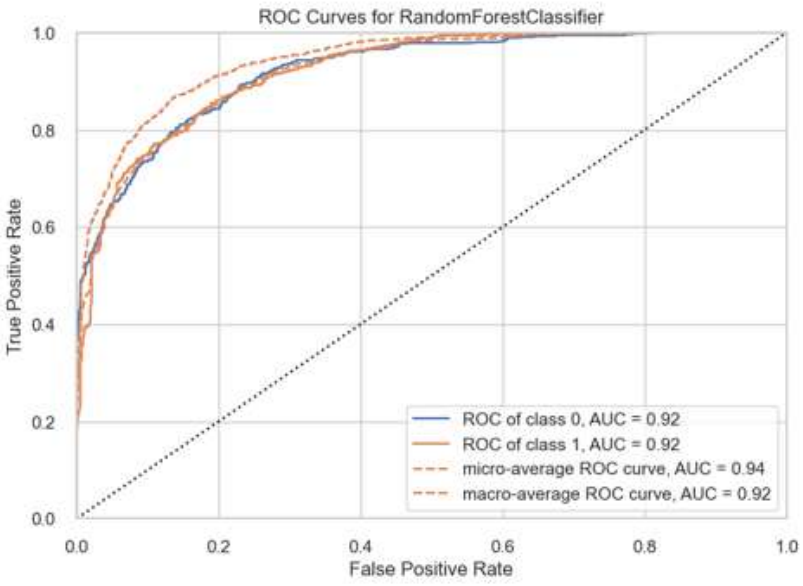
The ROC curves for the KNeighborsClassifier indicate good predictive ability with AUC scores of 0.91 for both classes. The micro-average and macro-average AUC scores are consistent at 0.93 and 0.91, respectively, suggesting balanced classification performance across classes.

SVM



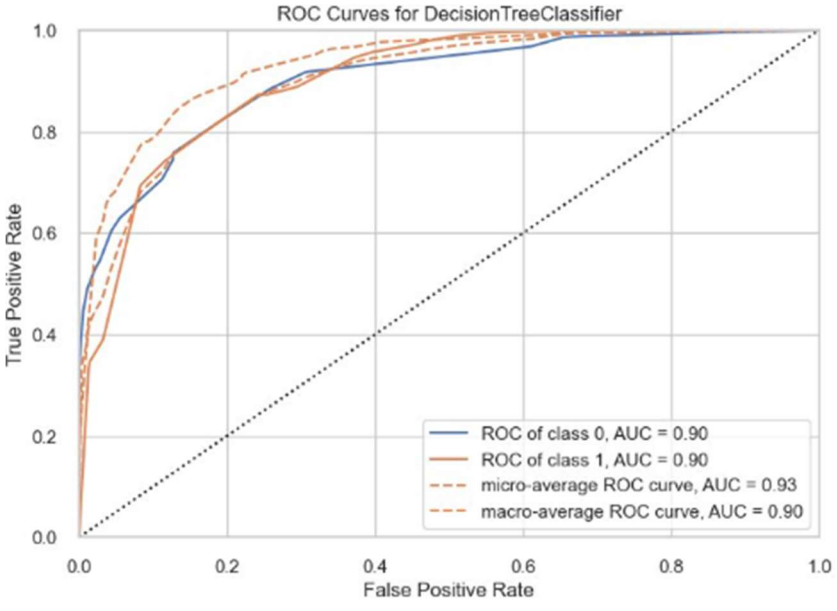
The ROC curve for the Support Vector Classifier (SVC) shows a high AUC of 0.93, indicating a strong ability to distinguish between the positive and negative classes. This suggests that the SVC model has a high true positive rate and a low false positive rate for the binary decision task.

Random Forest Classifier:



The ROC curves for the Random Forest Classifier demonstrate an excellent ability to classify both class 0 and class 1 with AUC values of 0.92. The micro-average AUC of 0.94 indicates outstanding overall performance, while the macro-average maintains a high AUC of 0.92, showing the model's consistent accuracy across classes.

Decision Tree Classifier:



The ROC curves for the DecisionTreeClassifier show good classification ability, with both class 0 and class 1 having an AUC of 0.90. The micro-average AUC of 0.93 suggests strong aggregate performance, while the macro-average AUC matches the class-specific AUCs, indicating balanced performance for both classes.

Combined Results:

	Algorithm	ROC AUC	Accuracy	Precision	F1 Scores
0	Logistic Regression	92.21	86.04	89.26	90.36
4	Random Forest	91.97	86.75	88.56	90.92
1	Kernel SVM	91.40	85.88	88.26	90.38
3	Gaussian NB	88.10	80.38	90.93	85.48
2	KNN	87.10	82.74	87.41	88.03
5	Decision Tree Classifier	77.27	81.50	87.26	86.80

The table ranks machine learning models by their performance metrics. Logistic Regression tops the list with the highest ROC AUC and strong overall metrics. Random Forest follows closely, with the highest accuracy. Gaussian NB has the best precision, while the Decision Tree Classifier trails with lower scores across all metrics.

Project Results

The primary aim of this machine learning project was to unearth the key factors that drive customer churn in the telecommunications industry. The initial phase focused on meticulously cleansing the dataset to eliminate any errors, inconsistencies, or missing data that might compromise the integrity of the analysis.

Following data preprocessing, we deployed a suite of machine learning algorithms, including Neural Networks, Logistic Regression, Support Vector Machine (SVM), and Random Forest, to model and predict customer churn. Our objective was to ascertain the influence of various parameters, such as billing patterns, service usage, customer demographics, and overall satisfaction, on a customer's likelihood to discontinue services.

The effectiveness of these models was assessed using metrics such as accuracy, F1-score, sensitivity, and specificity. We subjected each model to rigorous training and evaluation processes to establish their predictive prowess in terms of customer churn.

The analysis revealed that the Random Forest algorithm outperformed its counterparts in predicting customer churn. This superior performance is attributed to its resilience against outliers and its robustness in handling large datasets like ours. On the other hand, Logistic Regression's underperformance was linked to its assumption of linearity between dependent and independent variables. Similarly, SVM's diminished effectiveness was ascribed to the extensive size of our dataset, given that SVM generally performs less optimally with larger data volumes.

Impact of Project Outcomes

The implications of these findings are substantial for both the telecommunications company and its customers. For customers, this project paves the way for more tailored and satisfying service experiences. Customers can expect improvements in service offerings, more personalized engagement, and potentially more competitive pricing as the company leverages data-driven strategies to reduce churn.

For the company, the project's outcomes are pivotal in enhancing customer retention and cutting down the costs associated with customer turnover. By implementing proactive measures derived from the analysis, the company can maintain a loyal customer base, lower the expenses tied to acquiring new customers, and boost overall revenue. Additionally, the project provides valuable insights for identifying and rectifying potential issues in service delivery and customer relationship management, which, if unaddressed, could lead to increased churn rates.

The developed machine learning models offer predictive insights that can be instrumental for the company in forecasting and mitigating customer churn. These insights enable the company to invest in customer-centric initiatives, enhance service quality, and create an overall more engaging customer experience.

In essence, the impact of this project extends beyond mere churn prediction. It contributes to a more customer-focused, responsive, and efficient operational model for the telecommunications company, leading to a mutually beneficial relationship between the provider and its customers.