Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Customer Churn Analysis Report

Introduction:

In this report, we will be analyzing a customer churn dataset using the techniques we've learned in the course thus far. First we will use what we learned in Exploratory Data Analysis to learn about the dataset, then we will use different Regression and Classification techniques and compare their performance.

Data:

```
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   ID                      3150 non-null   int64
 1   Call Failure            3150 non-null   int64
 2   Complains               3150 non-null   object
 3   Subscription Length     3150 non-null   int64
 4   Charge Amount           3150 non-null   int64
 5   Seconds of Use          3150 non-null   int64
 6   Freq. of use            3150 non-null   int64
 7   Freq. of SMS            3150 non-null   int64
 8   Distinct Called Numbers 3150 non-null   int64
 9   Age Group               3150 non-null   int64
 10  Plan                    3150 non-null   object
 11  Status                  3150 non-null   object
 12  Age                     3150 non-null   int64
 13  Customer Value          3150 non-null   float64
 14  Churn                   3150 non-null   object
```

Musab Abuasi Student Id:1190954
Dr. Radi Jarrar

Details of Dataset:

ID: customer ID

Subscription Length: the duration of customer subscription (in months)
Freq. of use: the total number of calls
Freq. of SMS: the total number of text messages
Charge Amount: ordinal attribute in which 0 refers to the lowest amount and 9 refers to highest amount
Seconds of Use: total duration of calls in seconds
Distinct Numbers: total number of distinct phone calls
Call Failures: the total number of call failures
Complains: refers to if the customer have complains about the service or not
Age Group: ordinal attribute (1: younger age, 5: older age)
Age: the age of customer
Plan: prepaid or postpaid plan
Status: a binary attribute refers to the status of customers (active or not-active)
Churn: the class label (churn or non-churn)
Customer Value: a calculated value of customer (continuous attribute)

Exploratory Data Analysis:

Data Table:

	ID	Call Failure	Complains	Subscription Length	Charge Amount	Seconds of Use	Freq. of use	Freq. of SMS	Distinct Called Numbers	Age Group	Plan	Status	Age	Customer Value	Churn
0	1	3	no	10	2	1603	25	32	11	3	pre-paid	active	30	193.120	no
1	2	8	no	37	0	4255	65	0	13	2	pre-paid	active	25	194.400	yes
2	3	0	no	38	0	0	0	0	0	2	pre-paid	not-active	25	0.000	yes
3	4	10	no	36	0	2338	54	327	20	2	pre-paid	active	25	1579.140	yes
4	5	10	no	37	0	4083	60	0	31	1	pre-paid	active	15	227.865	yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3145	3146	0	no	16	0	1200	19	12	9	2	pre-paid	active	25	108.855	yes
3146	3147	9	no	15	0	5897	134	69	37	1	post-paid	active	15	711.205	yes
3147	3148	5	no	13	0	8437	164	57	35	4	pre-paid	active	45	357.525	yes
3148	3149	1	no	14	2	2357	38	15	14	3	pre-paid	active	30	155.800	yes
3149	3150	0	no	7	0	3895	33	199	5	3	pre-paid	active	30	953.120	yes

Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Statistics of Data:

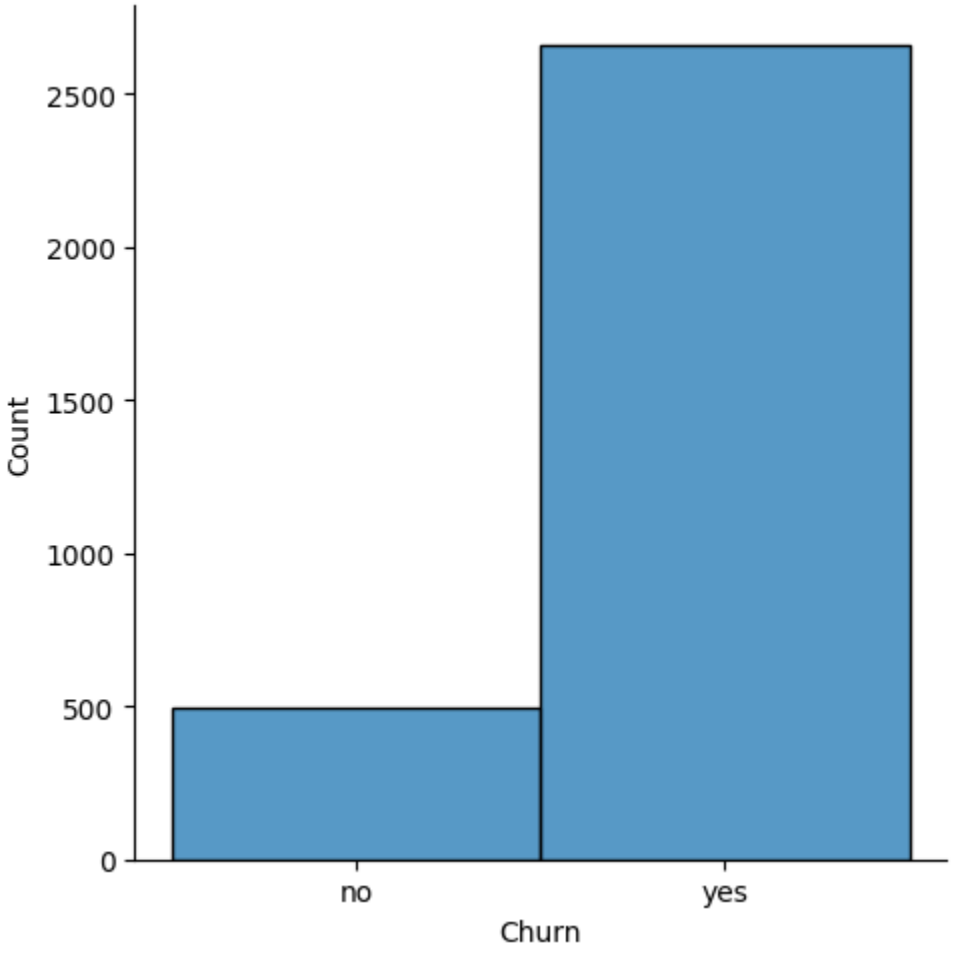
In the data table, we notice that there are 4 categorical columns: Plan,Status,Complains, and Churn. In order to keep everything numerical, we split these categories into two columns each instead.

	ID	Call Failure	Subscription Length	Charge Amount	Seconds of Use	Freq. of use	Freq. of SMS	Distinct Called Numbers	Age Group	Age	Customer Value
count	3150.000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000
mean	1575.500	7.627937	32.541905	0.942857	4472.459683	69.460635	73.174921	23.509841	2.826032	30.998413	470.972916
std	909.471	7.263886	8.573482	1.521072	4197.908687	57.413308	112.237560	17.217337	0.892555	8.831095	517.015433
min	1.000	0.000000	3.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	15.000000	0.000000
25%	788.250	1.000000	30.000000	0.000000	1391.250000	27.000000	6.000000	10.000000	2.000000	25.000000	113.801250
50%	1575.500	6.000000	35.000000	0.000000	2990.000000	54.000000	21.000000	21.000000	3.000000	30.000000	228.480000
75%	2362.750	12.000000	38.000000	1.000000	6478.250000	95.000000	87.000000	34.000000	3.000000	30.000000	788.388750
max	3150.000	36.000000	47.000000	10.000000	17090.000000	255.000000	522.000000	97.000000	5.000000	55.000000	2165.280000

Complains_no	Complains_yes	Plan_post-paid	Plan_pre-paid	Status_active	Status_not-active	Churn_no	Churn_yes
3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000
0.923492	0.076508	0.077778	0.922222	0.751746	0.248254	0.157143	0.842857
0.265851	0.265851	0.267864	0.267864	0.432069	0.432069	0.363993	0.363993
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	1.000000
1.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	1.000000
1.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	1.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

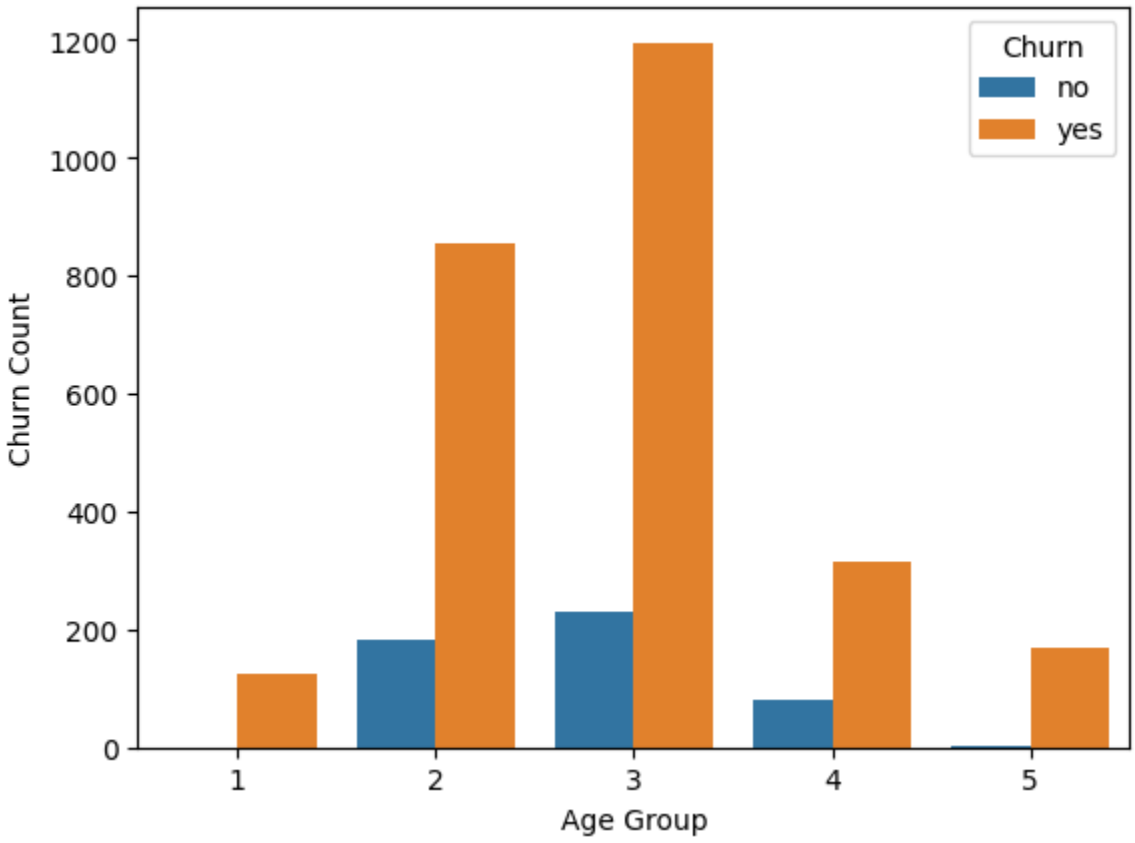
Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Distribution of Churn:



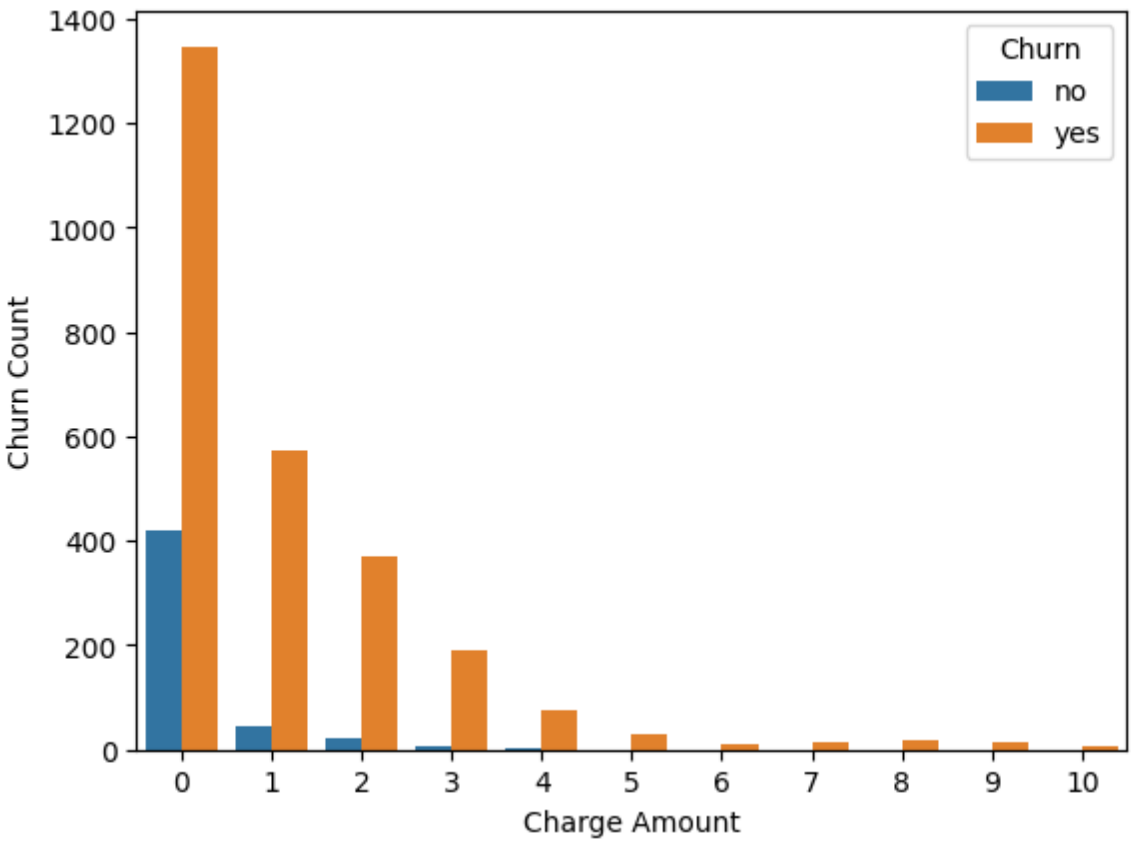
Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Churn Per Age Group:



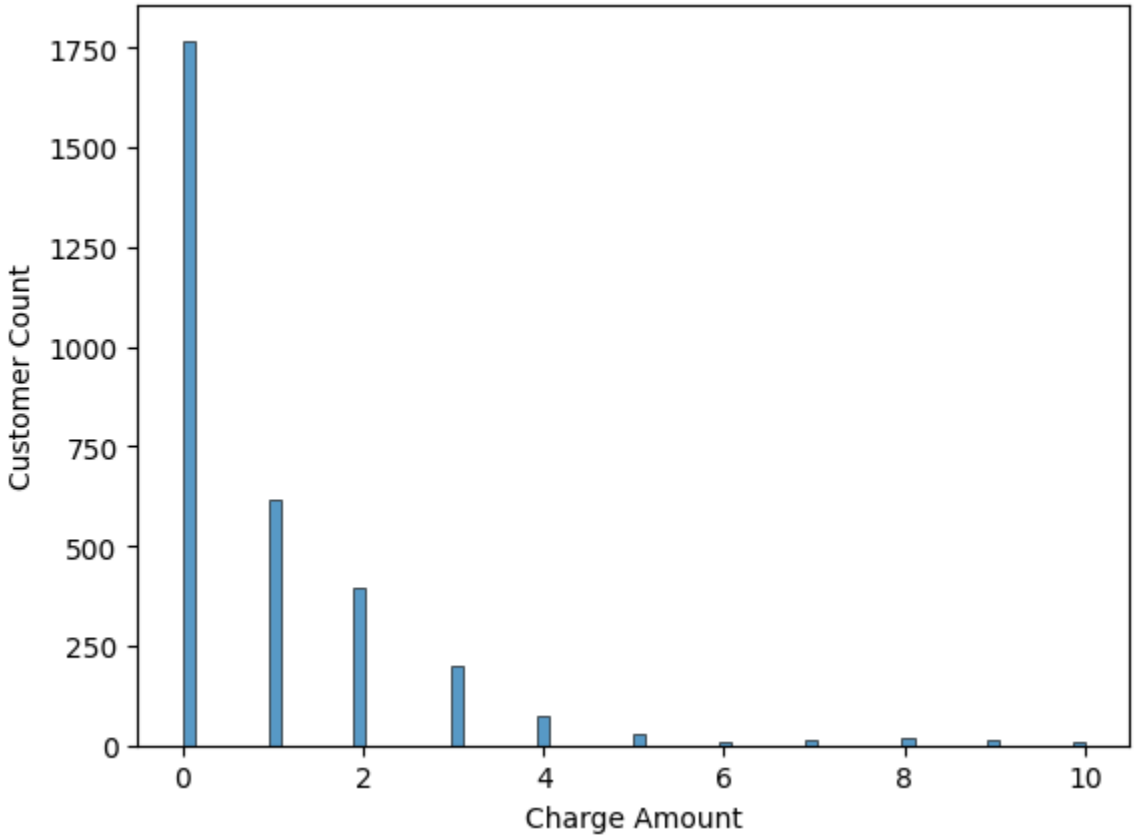
Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Churn Per Charge Amount:



Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Distribution and Statistics of Charge Amount:



	Charge Amount
count	3150.000000
mean	0.942857
std	1.521072
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	10.000000

Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Correlations:

	ID	Call Failure	Subscription Length	Charge Amount	Seconds of Use	Freq. of use	Freq. of SMS	Distinct Called Numbers	Age Group	Age	Customer Value	Complains _no	Complains _yes	Plan _post-paid	Plan _pre-paid
ID	1.00	-0.11	-0.05	-0.03	-0.02	-0.02	0.01	-0.00	-0.03	-0.02	0.01	0.00	-0.00	0.01	-0.01
Call Failure	-0.11	1.00	0.17	0.59	0.50	0.57	-0.02	0.50	0.05	0.04	0.12	-0.15	0.15	0.19	-0.19
Subscription Length	-0.05	0.17	1.00	0.08	0.12	0.11	0.08	0.09	0.02	-0.00	0.11	0.02	-0.02	-0.16	0.16
Charge Amount	-0.03	0.59	0.08	1.00	0.45	0.38	0.09	0.42	0.28	0.28	0.17	0.03	-0.03	0.32	-0.32
Seconds of Use	-0.02	0.50	0.12	0.45	1.00	0.95	0.10	0.68	0.02	0.02	0.42	0.10	-0.10	0.13	-0.13
Freq. of use	-0.02	0.57	0.11	0.38	0.95	1.00	0.10	0.74	-0.03	-0.03	0.40	0.09	-0.09	0.21	-0.21
Freq. of SMS	0.01	-0.02	0.08	0.09	0.10	0.10	1.00	0.08	-0.05	-0.09	0.92	0.11	-0.11	0.20	-0.20
Distinct Called Numbers	-0.00	0.50	0.09	0.42	0.68	0.74	0.08	1.00	0.02	0.05	0.28	0.06	-0.06	0.17	-0.17
Age Group	-0.03	0.05	0.02	0.28	0.02	-0.03	-0.05	0.02	1.00	0.96	-0.18	-0.02	0.02	-0.15	0.15
Age	-0.02	0.04	-0.00	0.28	0.02	-0.03	-0.09	0.05	0.96	1.00	-0.22	-0.00	0.00	-0.12	0.12
Customer Value	0.01	0.12	0.11	0.17	0.42	0.40	0.92	0.28	-0.18	-0.22	1.00	0.13	-0.13	0.25	-0.25
Complains _no	0.00	-0.15	0.02	0.03	0.10	0.09	0.11	0.06	-0.02	-0.00	0.13	1.00	-1.00	-0.00	0.00
Complains _yes	-0.00	0.15	-0.02	-0.03	-0.10	-0.09	-0.11	-0.06	0.02	0.00	-0.13	-1.00	1.00	0.00	-0.00
Plan _post-paid	0.01	0.19	-0.16	0.32	0.13	0.21	0.20	0.17	-0.15	-0.12	0.25	-0.00	0.00	1.00	-1.00
Plan _pre-paid	-0.01	-0.19	0.16	-0.32	-0.13	-0.21	-0.20	-0.17	0.15	0.12	-0.25	0.00	-0.00	-1.00	1.00
Status _active	-0.01	0.11	-0.14	0.36	0.46	0.45	0.30	0.41	-0.00	0.00	0.41	0.27	-0.27	0.16	-0.16
Status _not-active	0.01	-0.11	0.14	-0.36	-0.46	-0.45	-0.30	-0.41	0.00	-0.00	-0.41	-0.27	0.27	-0.16	0.16
Churn _no	0.00	-0.01	-0.03	-0.20	-0.30	-0.30	-0.22	-0.28	-0.01	-0.02	-0.29	-0.53	0.53	-0.11	0.11
Churn _yes	-0.00	0.01	0.03	0.20	0.30	0.30	0.22	0.28	0.01	0.02	0.29	0.53	-0.53	0.11	-0.11

Musab Abuasi
Dr. Radi Jarrar

Student Id:1190954

Plan_post-paid	Plan_pre-paid	Status_active	Status_not-active	Churn_no	Churn_yes
0.01	-0.01	-0.01	0.01	0.00	-0.00
0.19	-0.19	0.11	-0.11	-0.01	0.01
-0.16	0.16	-0.14	0.14	-0.03	0.03
0.32	-0.32	0.36	-0.36	-0.20	0.20
0.13	-0.13	0.46	-0.46	-0.30	0.30
0.21	-0.21	0.45	-0.45	-0.30	0.30
0.20	-0.20	0.30	-0.30	-0.22	0.22
0.17	-0.17	0.41	-0.41	-0.28	0.28
-0.15	0.15	-0.00	0.00	-0.01	0.01
-0.12	0.12	0.00	-0.00	-0.02	0.02
0.25	-0.25	0.41	-0.41	-0.29	0.29
-0.00	0.00	0.27	-0.27	-0.53	0.53
0.00	-0.00	-0.27	0.27	0.53	-0.53
1.00	-1.00	0.16	-0.16	-0.11	0.11
-1.00	1.00	-0.16	0.16	0.11	-0.11
0.16	-0.16	1.00	-1.00	-0.50	0.50
-0.16	0.16	-1.00	1.00	0.50	-0.50
-0.11	0.11	-0.50	0.50	1.00	-1.00
0.11	-0.11	0.50	-0.50	-1.00	1.00

(table cont.)

Features that have High Correlation:

1. Seconds of Use and Freq. of Use (0.95 , High Correlation)
    - The duration of each call is highly correlated to the total amount of calls received because the more often something is used, the more time it is in operation, and the longer it is in operation, the more often it is used
2. Seconds of Use and Distinct Called Numbers (0.68, High Correlation)
    - The duration of each call is highly correlated to the total amount of distinct calls received
3. Freq of Use and Distinct Called Numbers (0.74, High Correlation)
    - the more time a device is in use, the more likely it is to have made contact with a large number of distinct called number

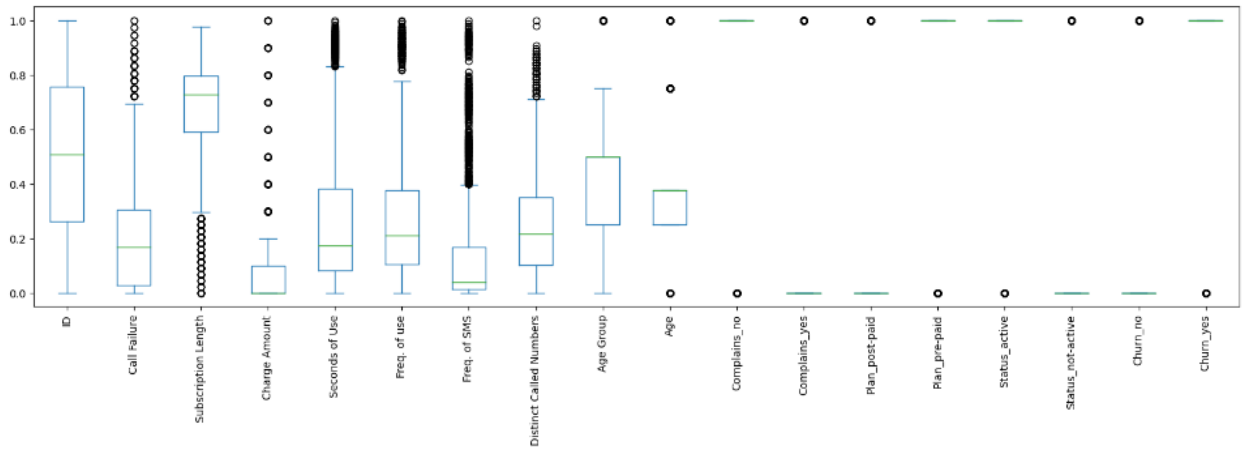
4. Freq of SMS and Customer Value (0.92, High Correlation)
    - The more messages the customer is receiving the more valuable he probably is
5. Age Group and Age (0.96, High Correlation)
    - Obviously the higher the age group then the higher the age and vice versa

Preprocessing:

We need to feature scale the dataset so we can have all the columns in the same range. Here we fit all columns between 0 and 1.

Dataset after Feature Scaling:

BoxPlot:



Statistics Summary:

	ID	Call Failure	Subscription Length	Charge Amount	Seconds of Use	Freq. of use	Freq. of SMS	Distinct Called Numbers	Age Group	Age	Customer Value
count	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000
mean	0.500000	0.211887	0.671407	0.094286	0.261700	0.272395	0.140182	0.242369	0.456508	0.399960	0.217511
std	0.288813	0.201775	0.194852	0.152107	0.245635	0.225150	0.215014	0.177498	0.223139	0.220777	0.238775
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.250000	0.027778	0.613636	0.000000	0.081407	0.105882	0.011494	0.103093	0.250000	0.250000	0.052557
50%	0.500000	0.166667	0.727273	0.000000	0.174956	0.211765	0.040230	0.216495	0.500000	0.375000	0.105520
75%	0.750000	0.333333	0.795455	0.100000	0.379067	0.372549	0.166667	0.350515	0.500000	0.375000	0.364105
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Complains_no	Complains_yes	Plan_post-paid	Plan_pre-paid	Status_active	Status_not-active	Churn_no	Churn_yes
3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000	3150.000000
0.923492	0.076508	0.077778	0.922222	0.751746	0.248254	0.157143	0.842857
0.265851	0.265851	0.267864	0.267864	0.432069	0.432069	0.363993	0.363993
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	1.000000
1.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	1.000000
1.000000	0.000000	0.000000	1.000000	1.000000	0.000000	0.000000	1.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Linear Regression Modeling:

1. LRM1

Apply linear regression to learn the attribute “Customer Value” using all independent attributes (call this model LRM1).

```
regr = linear_model.LinearRegression()
LRM1=regr.fit(X_train, y_train)
y_pred_LRM1 = regr.predict(X_test)
```

2.LRM2

Apply linear regression using the set of 3-most important features (from your point of view); and explain why did you use these 3 attributes (call this model LRM2).
In my point of view I believe that Subscription Length, Freq. of use and activity affect the Customer value. If he is subscribed for a long time then it makes sense this customer is more valued then someone that is new. If a customer is using the product more frequently then he probably is valued. Finally if a customer is active then it probably means he is engaged with the product.

```
LRM2=regr.fit(X_train[['Subscription Length', 'Freq. of use', 'Status_active']], y_train)
y_pred_LRM2 = regr.predict(X_test[['Subscription Length', 'Freq. of use', 'Status_active']])
```

Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

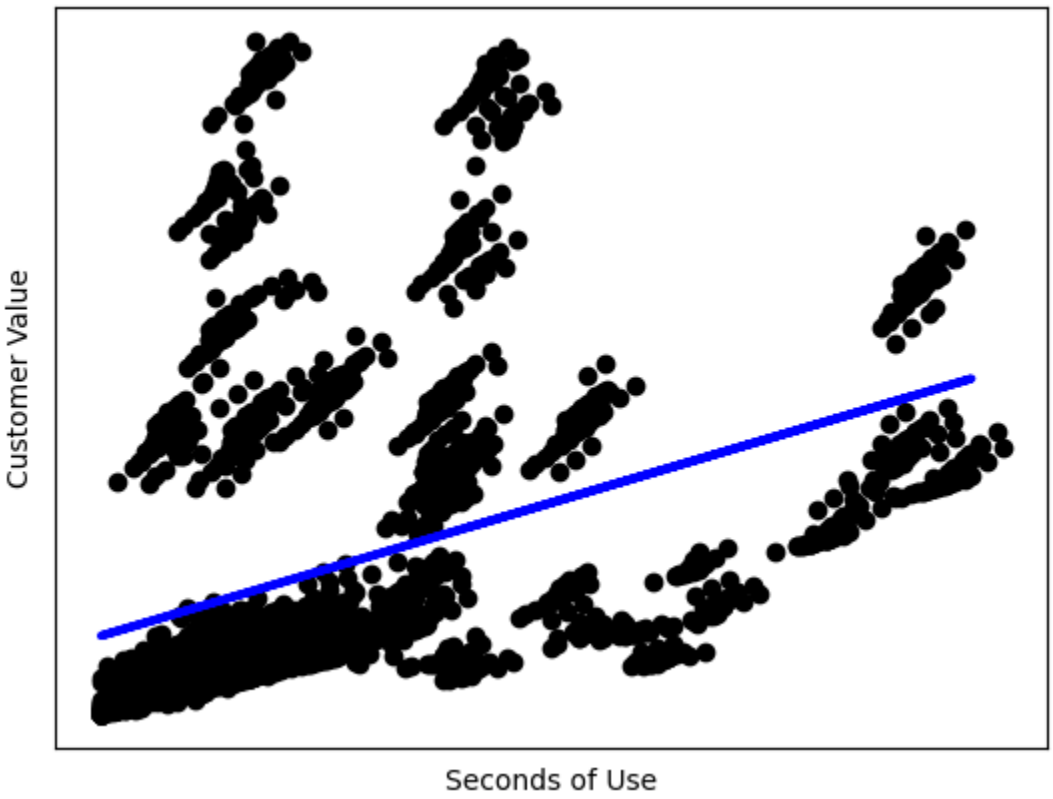
3. LRM3

Apply linear regression using the set of the most important features (based on the correlation coefficient matrix) and explain why did you use these 3 attributes (call this model LRM3).

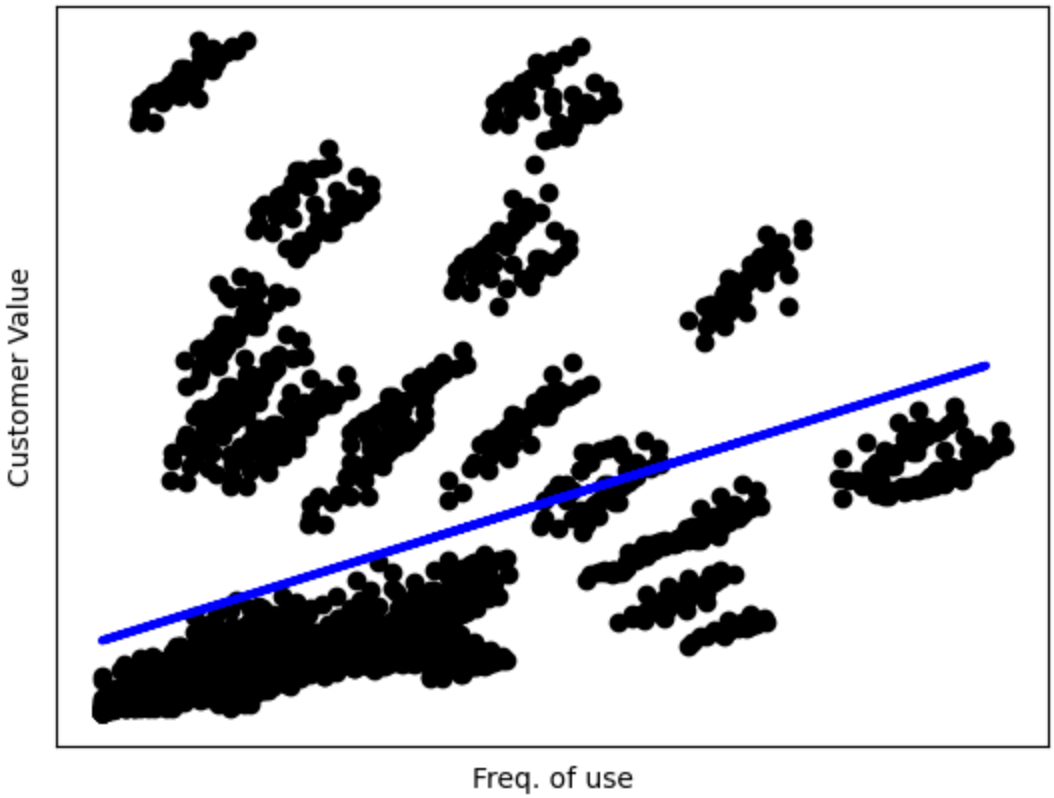
Freq. of SMS, Freq. of Use, and Seconds of use show the most correlation with customer value. Freq of SMS(0.92), Freq. of Use (0.42), Seconds of Use(0.41)

```
LRM3=regr.fit(X_train[['Freq. of SMS','Freq. of use','Seconds of Use']], y_train)
y_pred_LRM3 = regr.predict(X_test[['Freq. of SMS','Freq. of use','Seconds of Use']])
```

Plotting the attributes using Linear Regression:



Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954



Linear Regression Results:

We will be comparing the models based on the mean squared error(shows the error between actual values and the predicted values), the mean absolute error (shows the error between actual values and the predicted values using absolute value) and the R^2 Score (how well does the line fit).

	Mean Squared Error	Mean Absolute Error	R^2 Score
LRM1	0.001041	0.018545	0.981533
LRM2	0.041871	0.151001	0.257295
LRM3	0.002118	0.024388	0.962427

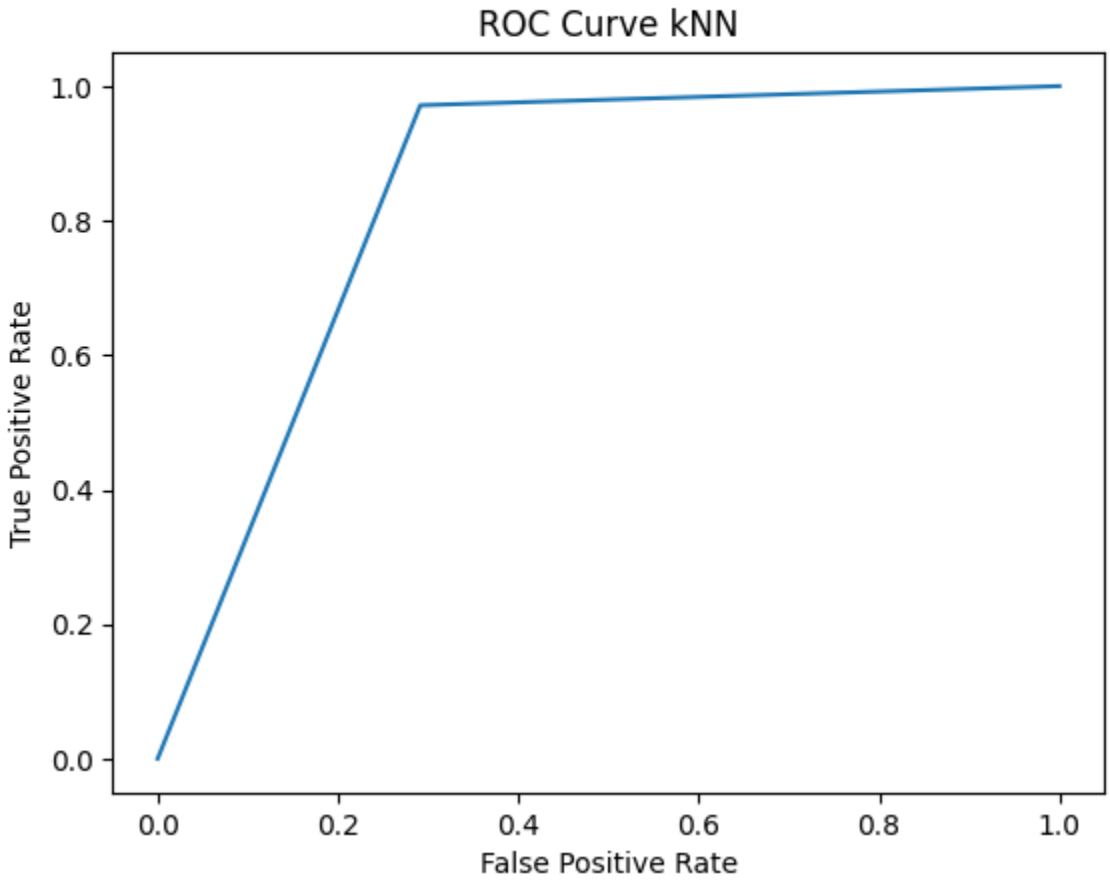
Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Classification Modeling:

1. kNN

Run k-Nearest Neighbours classifier to predict churn of customers (the “Churn” feature) using the test set

	Predicted Negative	Predicted Positive
Actual Negative	73	30
Actual Positive	15	512

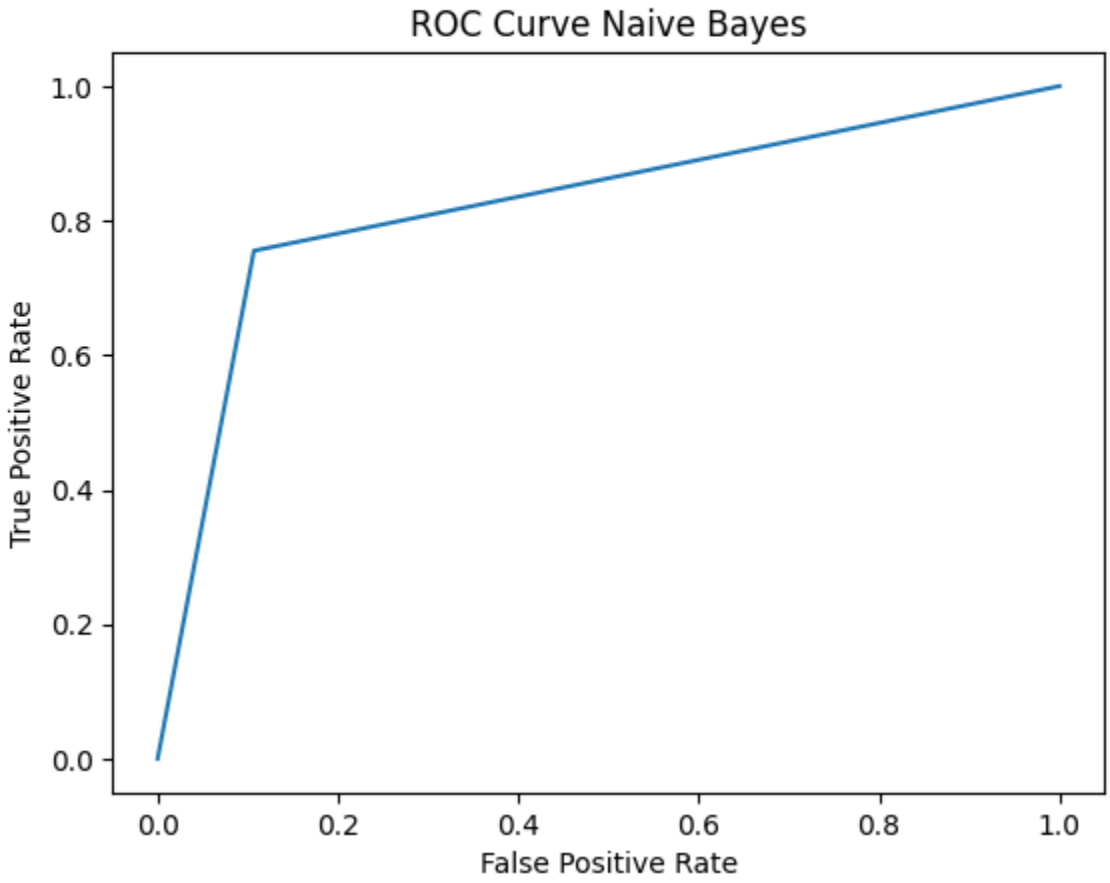


Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Naive Bayes

Run Naive Bayes classifier to predict churn of customers (the “Churn” feature) using the test set

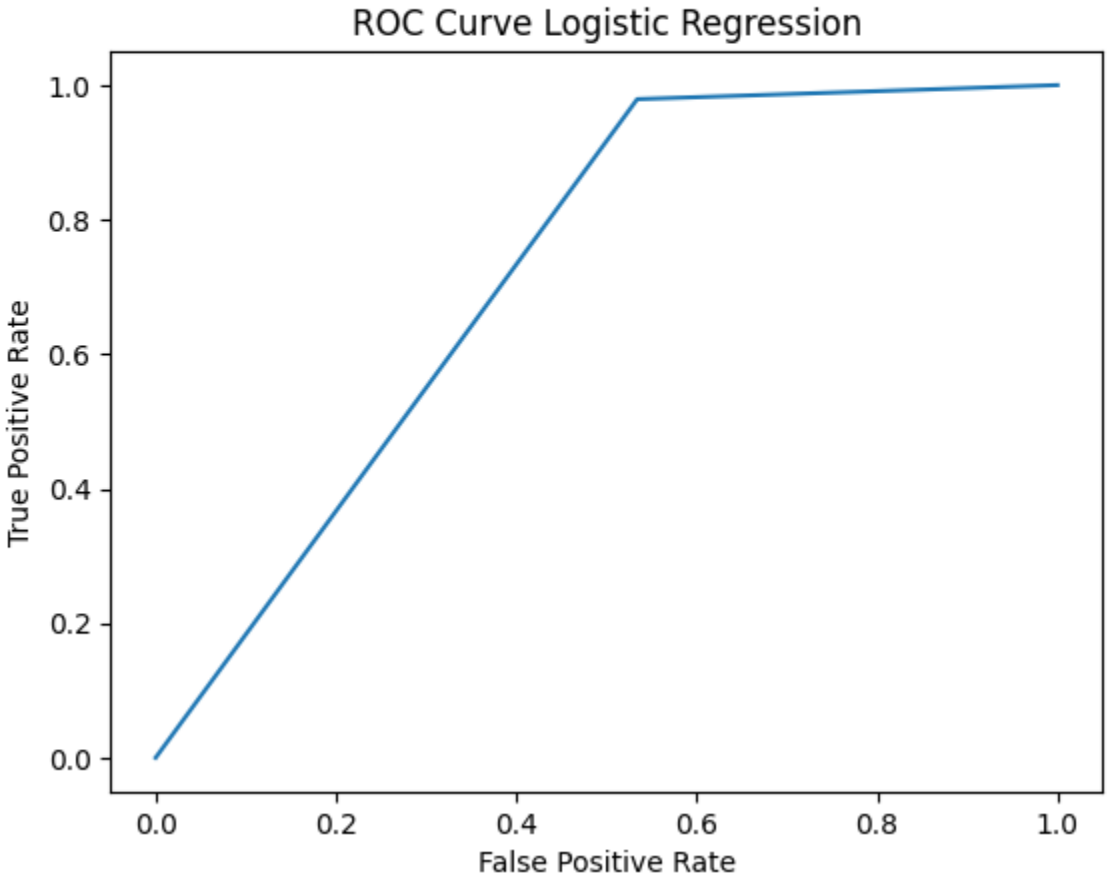
	Predicted Negative	Predicted Positive
Actual Negative	92	11
Actual Positive	129	398



Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Logistic Regression:

	Predicted Negative	Predicted Positive
Actual Negative	48	55
Actual Positive	11	516



Musab Abuasi
Dr. Radi Jarrar
Student Id:1190954

Classification Results:

	ROC_AUC_SCORE
kNN	0.840137
Naive Bayes	0.824211
Logistic Regression	0.722573

ROC_AUC_SCORE column shows how well the model can distinguish between negative and positive classes. Here we see kNN has the highest score, and this is probably due to the fact that it manages to have the highest peak towards true positive and a low false positive value (x=0.3,y=0.9) which gives it more area, hence why it scores better than Naive Bayes and Logistic Regression.