BIRZEIT UNIVERSITY

Computer Science Department
Dr. Radi Jarrar
MACHINE LEARNING COMP4388
Section: 1 T, R.

The Customer Churn Dataset

Telecommunication Company

Project one

15. December. 2022

Dear Telecommunication company representative,

After completing the analysis of the Customer Churn Data you provided, I'm writing to share the results. Based on the following analysis you will find conclusions that will better your decision making concerning the reducing of your customer churn and improving you customer's value, with emphasizing such a step to maintain your customer retention.

With respect,

Sondos Aabed,
1190652

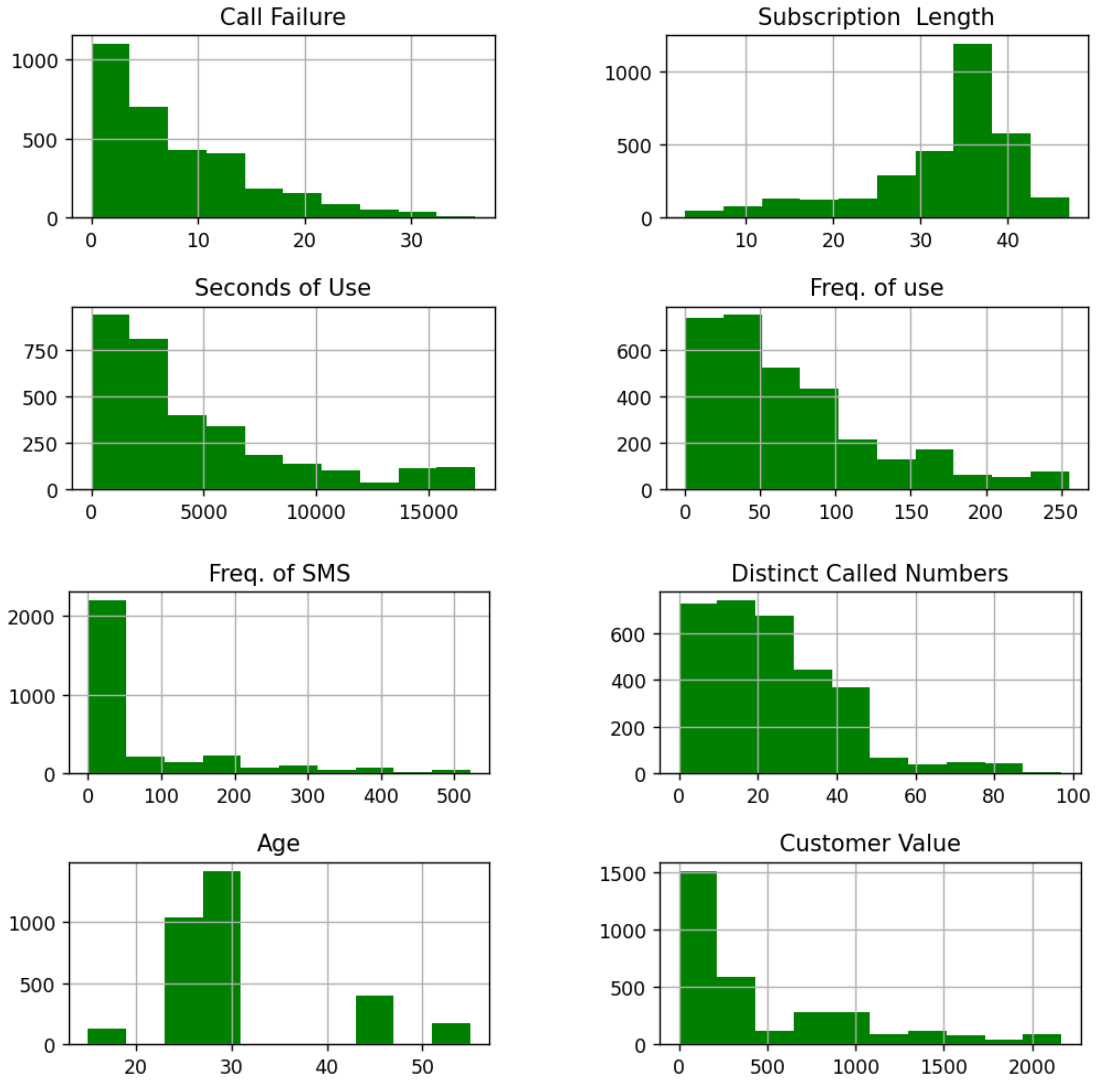
Contents

Summary Statics

In this section, the summary statistics of all attributes in the dataset is presented. The dataset includes 3150 observations and 15 attributes. The attributes are divided into two categories: quantitative and qualitative, hence for each there is a proper summary statics provided. ***(Task 1)***

Quantitative attributes

These attributes include age, customer value, call failures, seconds of use, distinct numbers, frequency of SMS, frequency of use, and subscription length. In order to visualize the distribution of the quantitative attributes, univariate histograms were used as follow:



It is seen that both call failure and seconds of use histograms are skewed to the left.

In addition to histograms, the table below, shows useful statics for each quantitative attribute that includes the following: mean, standard deviation, minimum, quartiles and maximum.

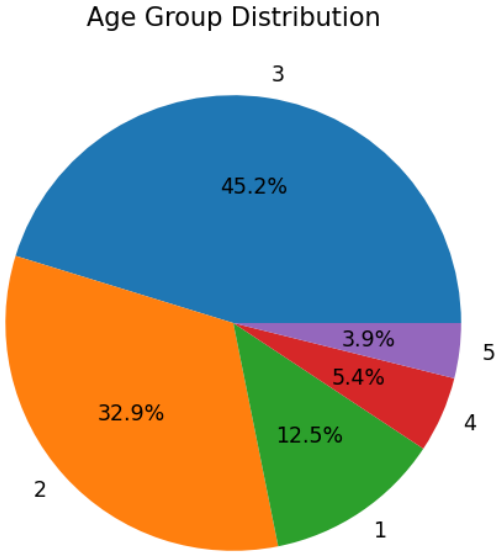
	Mean	***Standard deviation***	***Minimum***	***25%***	***50%***	***75%***	***Maximum***
Age	*30.998*	*8.831*	*15*	*25*	*30*	*30*	*55*
Customer Value	*470.9723*	*517.015*	*0.000*	*113.801*	*228.480*	*788.389*	*2165.280*
Call Failure	*7.628*	*7.264*	*0.000*	*1.000*	*6.000*	*12.000*	*36.000*
Subscription Length	*32.542*	*8.573*	*3.000*	*30.000*	*35.000*	*38.000*	*47.000*
Seconds Of use	*4472.451*	*4197.909*	*0.000*	*1391.250*	*2990.000*	*6478.250*	*17090.000*
Frequency of Calls	*69.461*	*57.4133*	*0.000*	*27.000*	*54.000*	*95.000*	*255.000*
Frequency of SMS	*69.461*	*57.413*	*0.000*	*27.000*	*54.000*	*95.000*	*255.000*
Distinct numbers	*23.501*	*17.217*	*0.000*	*10.000*	*21.000*	*34.000*	*97.000*

Overall, the age of your customers ranges from 15 to 55 years old, with a mean of 31 years. These customers have a number of calls failures that ranges from 0 failed call to 36 failed call, with a mean of 7.628 failed call for each customer.

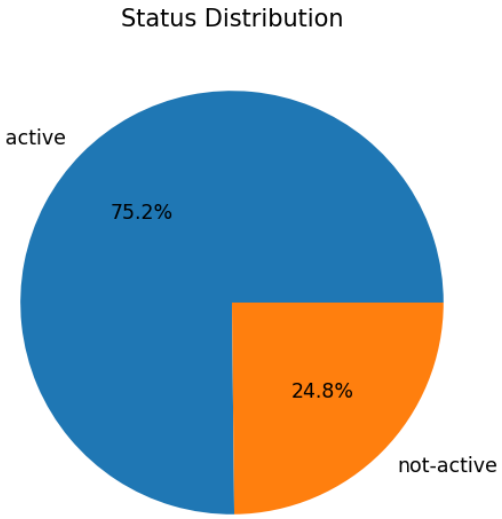
Qualitative attributes

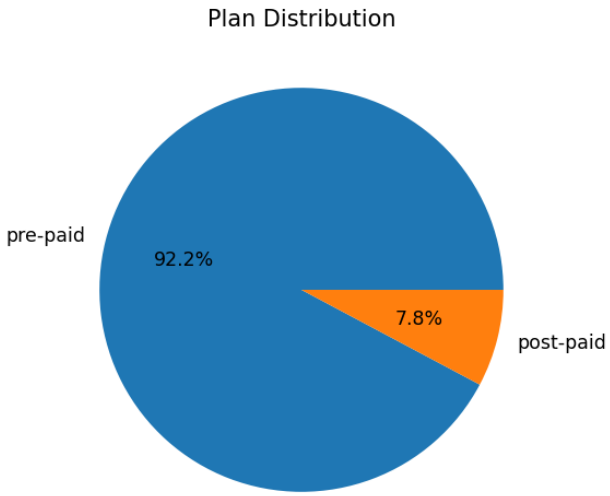
These attributes include: age group, charge amount, churn, status, plan, and complains. In order to visualize the distribution of the qualitative attributes, univariate pie and bar charts were used.

The following chart illustrates the distribution of age group of your churn customers: it shows that 45.2% of them are from Age Group 3, meaning that age group 3 is dominating.

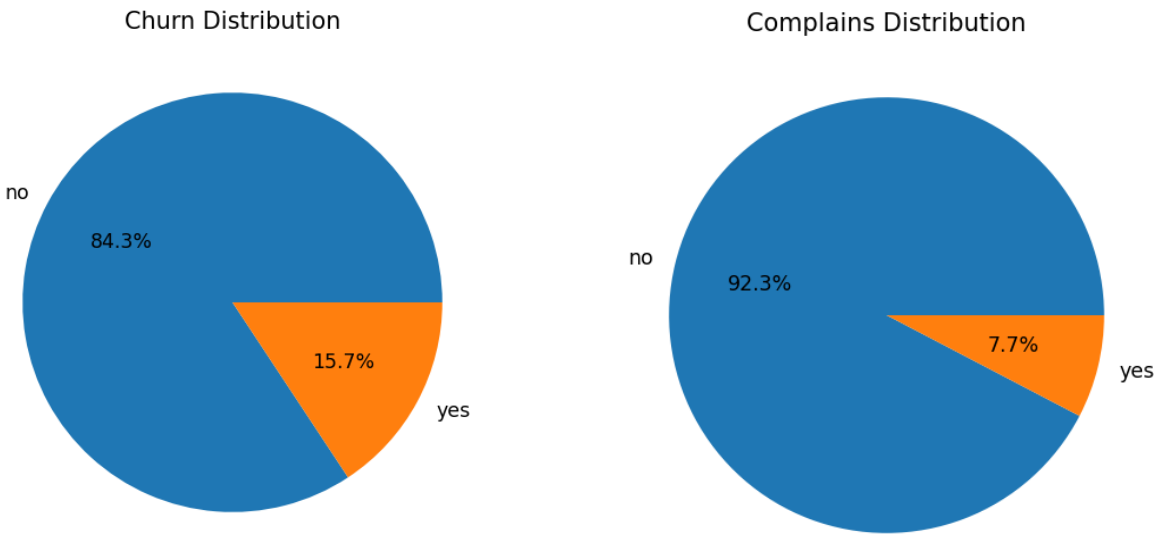


Here, the first pie chart shows that a majority of 75.2% of the customers are currently active, while 24.8% are non-active. The second chart shows that 92.2% of the customers preferred pre-paid plan to 7.8% who preferred post-paid.

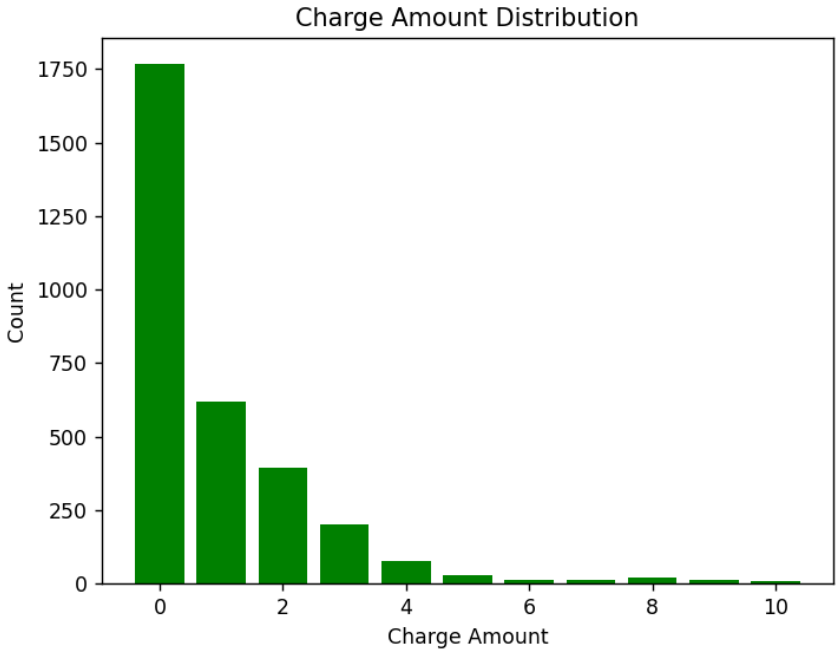




(Task 2) *Based on the following pie charts, about 15.7% of the customers have churned, which is relatively not quite a lot. While a smaller portion, about 7.7% of all customers have made complains the majority did not make any complains at all.*



(Task 5) *The bellow chart shows the details of charge amount that you charge the customers. In which it shows that 50% of them are zero charged for their use:*

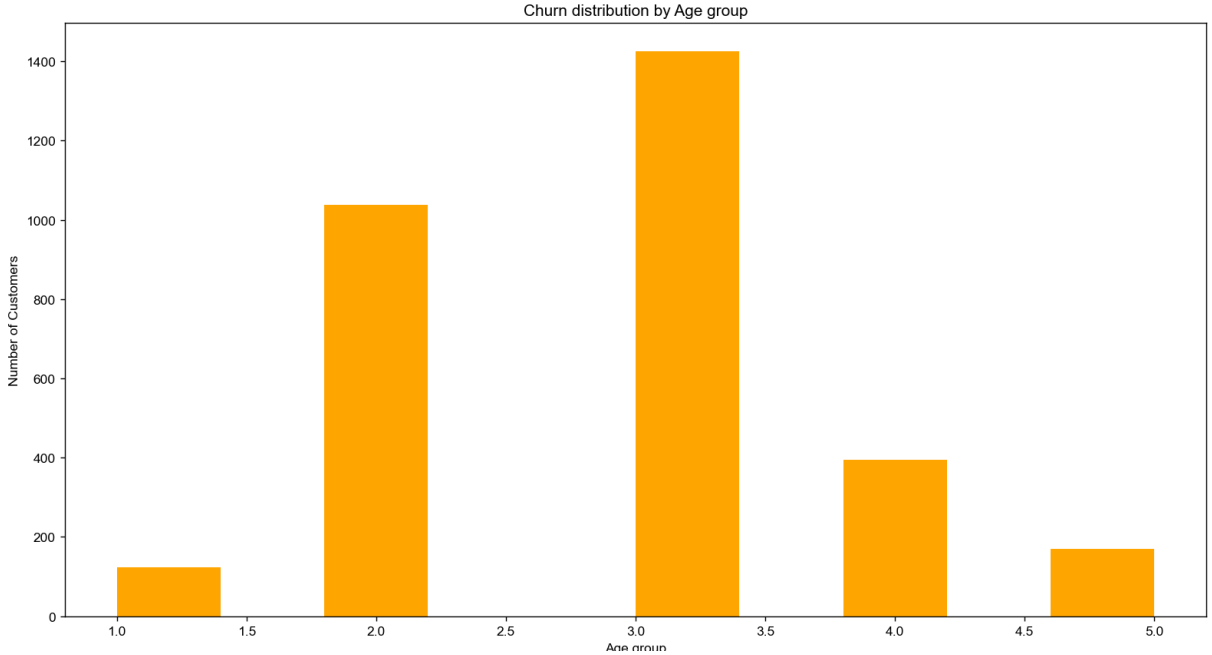


Features Scaling

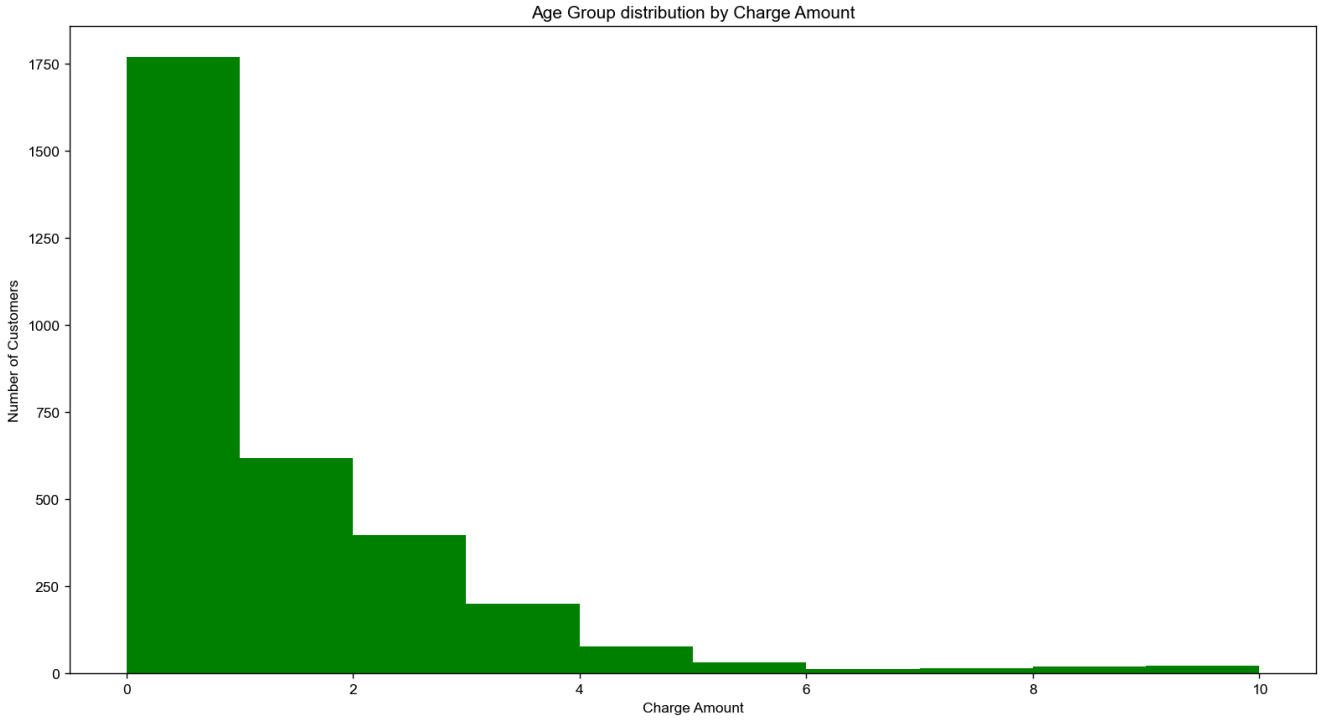
The min-max feature scaling was performed on the dataset.

Multivariate analysis

(Task 3) *The following histogram details the amount of churn in each sub-group for each age group. Which shows dominance of age group 3.*

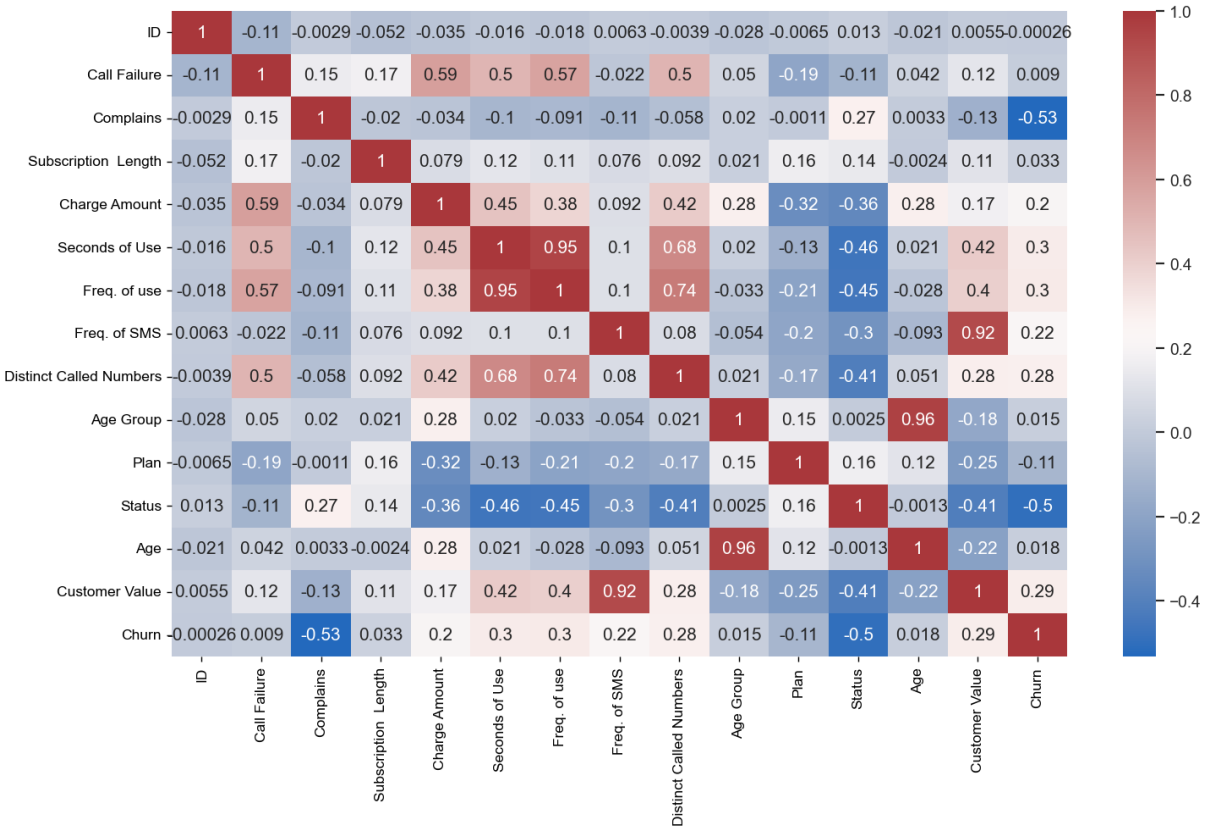


(Task 4) *The following histogram details the amount of churn in each sub-group for each charge amount.*



Data Correlation ***(Task 6)***

Now that the statics summary of all the dataset attributes is presented, the correlation between all the attributes is visualized using the Heat map which represents Peterson's coefficient between all the pairs of attributes. In order to do so, the qualitative attributes were converted into quantitative attributes as follows:



Because a column is perfectly correlated with itself, the diagonal values are all equal to 1. However, there is high correlation between Age and Age group given the attributes are related, the decision is to keep the Age attribute since they are correlated with 0.96 and drop the Age Group.

For the Customer Value it is highly correlated with Frequency of SMS. The value of the customer is predicted using the frequency of the SMS with a Pearson correlation coefficient of 0.92. It is also noticed that there is a high correlation between Frequency of use and Seconds of calls, which makes sense, they are related.

The attribute Complains is positively correlated with Status, with a Pearson correlation coefficient of 0.271. It is noticed to be negatively correlated with Call Failure, and Churn. Which potentially indicates that they are nonlinearly correlated, since when interpreting the results one could think that complains and Churn would positively correlate.

The ID attribute will be dropped because for all attributes it is negatively correlated. Nonetheless, for all the dropped attributes, their correlation does not imply causation in this case it means they could be predicted from the other kept attribute.

It is noticed that using the Heat Map was not enough at this point of analysis to decide which attributes to drop or select. However, these only two attributes will be ***dropped****:*

1- *ID*
2- *Age Group*

These twelve attributes will be ***kept*** *for the learning process:*

1- *Churn*
2- *Age*
3- *Call Failure*
4- *Complains*
5- *Subscription Length*
6- *Charge Amount*
7- *Frequency of Use*
8- *Customer Value*
9- *Distinct calls numbers*
10- *Seconds of Calls*
11- *Frequency of SMS*
12- *Plan*

Now that the attributes are selected, they will be cleaned. This involves, removing duplicate entries, filling in missing values with the mean of the attribute and standardizing formats. So that the learning models will get cleansed dataset. Then it will be split into test and training sets, 20:80.

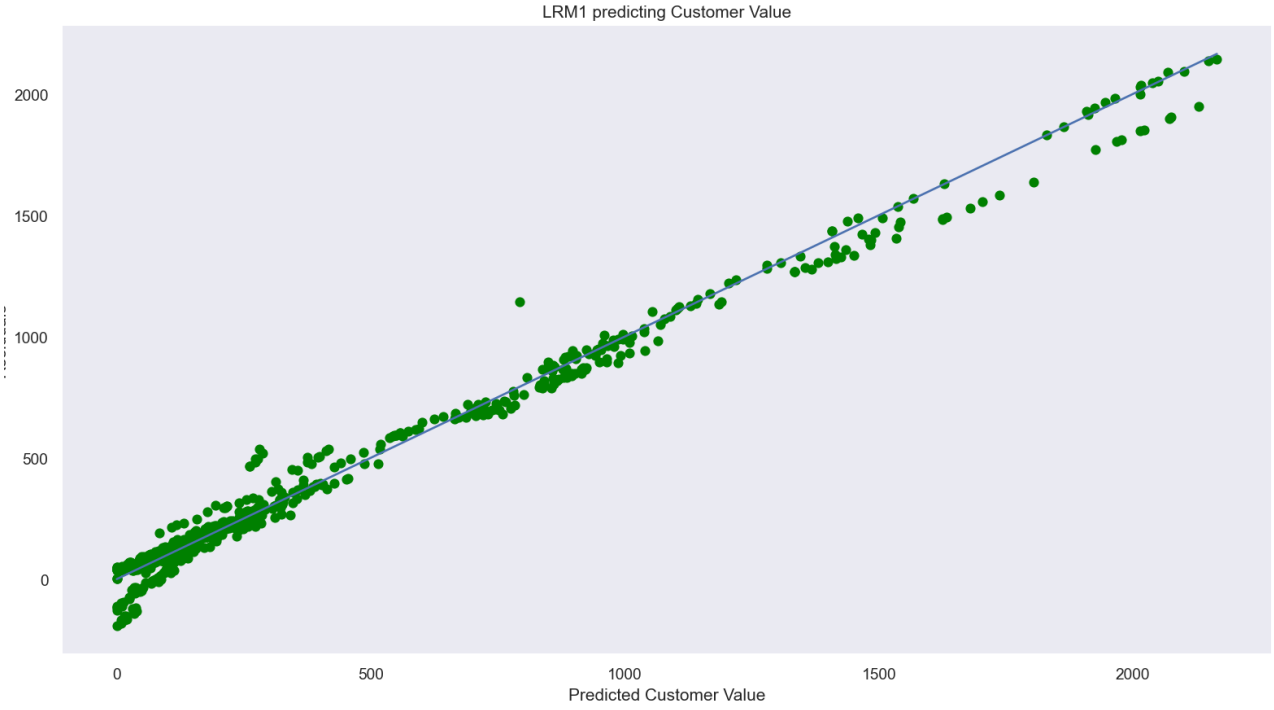
Linear Regression

In this section, the Linear Regression will be used in the Customer Churn dataset. At the end of this section a comparison of performance for the three learning models will be presented.

Customer Value LRM1

After applying linear regression to learn the attribute "Customer Value" and using all independent attributes, the R-Square value of the method is: 0.98.

The bellow chart shows the actual and the predicted values after applying linear regression:



The predicted data appears to follow the line, nonetheless there are a few outliers of the line. The linear regression model LRM1 appears to be good fitting.

The Mean Squared Error for this model was measured: 4502.59

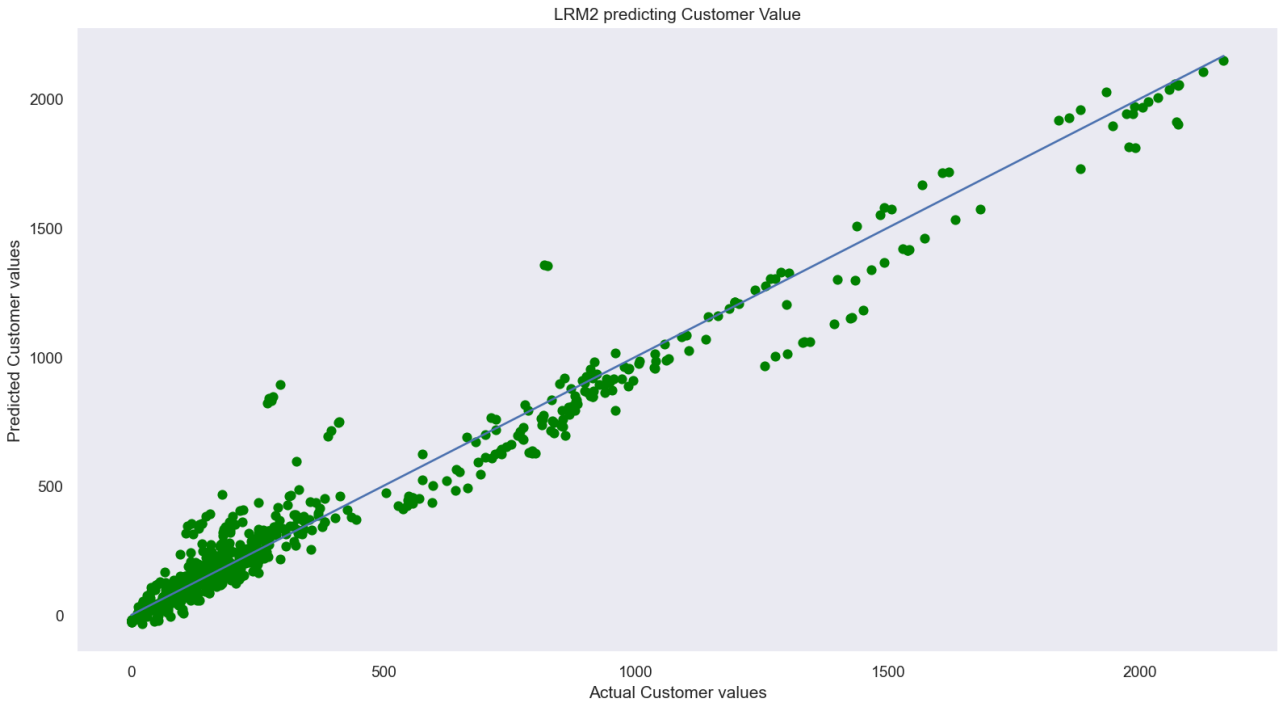
Point of view attributes LRM2

These following attributes were considered the most important since they are the most relevant to the Customer Value:

1- *Frequency of SMS*
2- *Frequency of use*
3- *Subscription length*

Both frequency of use and frequency of SMS indicates a level of engagement from a customer. For example if a customer has higher values of those attributes, that indicates that she has higher Customer value. As for the subscription length, it is seen as an indicator of commitment and paying for the services. Typically, these attributes would be most important to calculate a customer value in a telecommunication company.

The bellow chart shows the actual and the predicted values after applying linear regression to learn the attribute "Customer Value" using above attributes:



The linear regression model LRM2 appears to be fitting, because the data appears to follow the linear tend but there are few outliers.

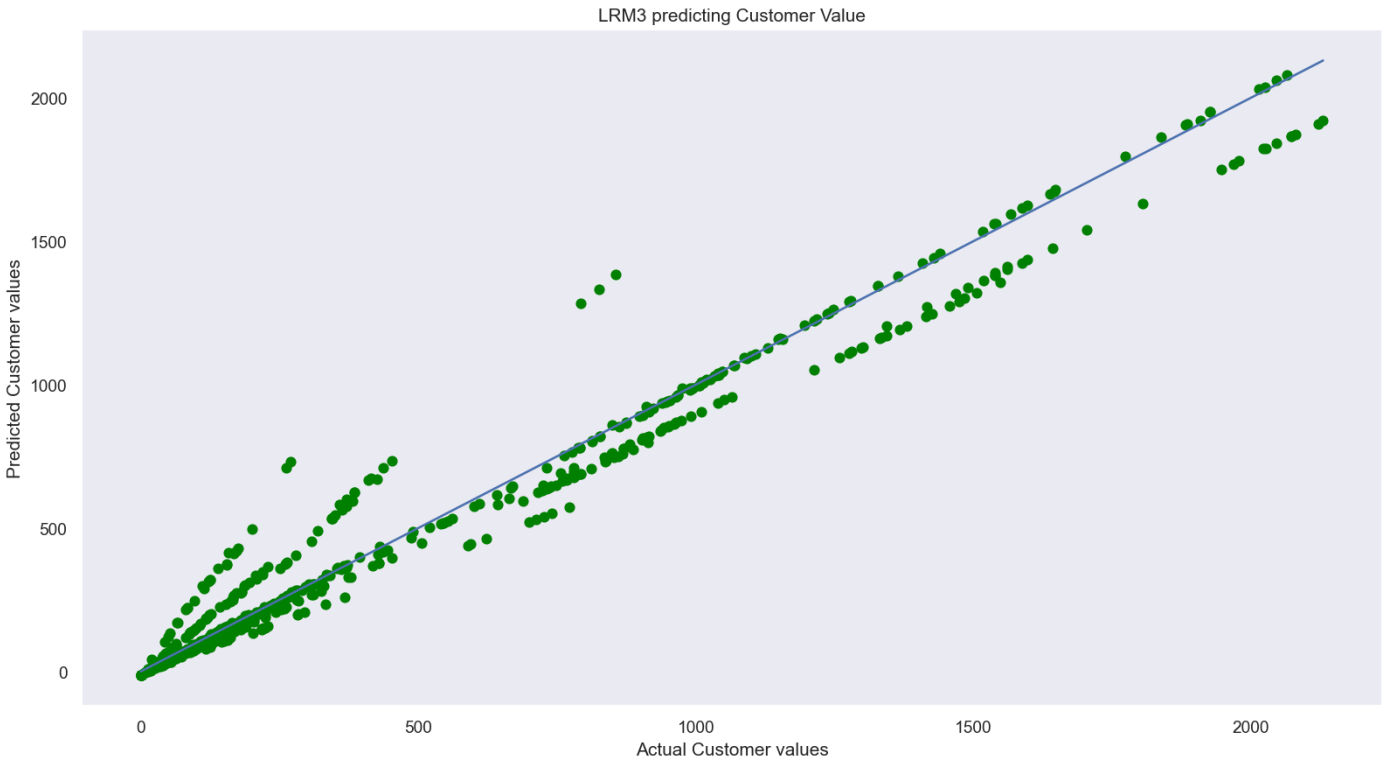
As for the Mean Squared Error for this model it was measured: 12359.32

Correlation matrix attributes LRM3

These following attributes were considered the most important as it was shown in section ***Data Correlation*** *because they had the higher correlation to the Customer Value based on the correlation coefficient matrix*

1- *Frequency of SMS*
2- *Frequency of use*
3- *Seconds of use*

The bellow chart shows the actual and the predicted values after applying linear regression to learn the attribute "Customer Value" using above attributes:



The linear regression model LRM3 appears to be fitting, because the data appears to follow the linear tend.

The Mean Squared Error for this model was measured: 8750.31

Comparison of Performance

In comparing the three linear regression models, it is clear in the scatter plots that they all performed differently in terms of prediction accuracy. An adequate performance metric for these linear regression models would be the Mean Squared error and R squared, shown in the table below:

	LRM1	***LRM2***	***LRM3***
MSE	*4502.59*	*12359.32*	*8750.31*
R^2	*0.98*	*0.95*	*0.97*

LRM1, which used all independent attributes, had a mean squared error of 4502.59 and 0.98 R^2. On the other hand LRM2 which used the most important three attributes: Subscription length, Frequency of use and SMS had a higher mean squared error of 10135.80. As for LRM3 which used the most important attributes according to the correlation matrix: Seconds of use and Frequency of use and SMS had a higher MSE 8750.31 and 0.97 R^2.

It appears that LRM1 and LRM3 may be the best models out of the three based on their MSE and R2 values. This result would point out the importance of selecting the most relevant and informative attributes for the model, as it can affect the performance, and it is not always how it appears to be more relevant to the target.

Classification

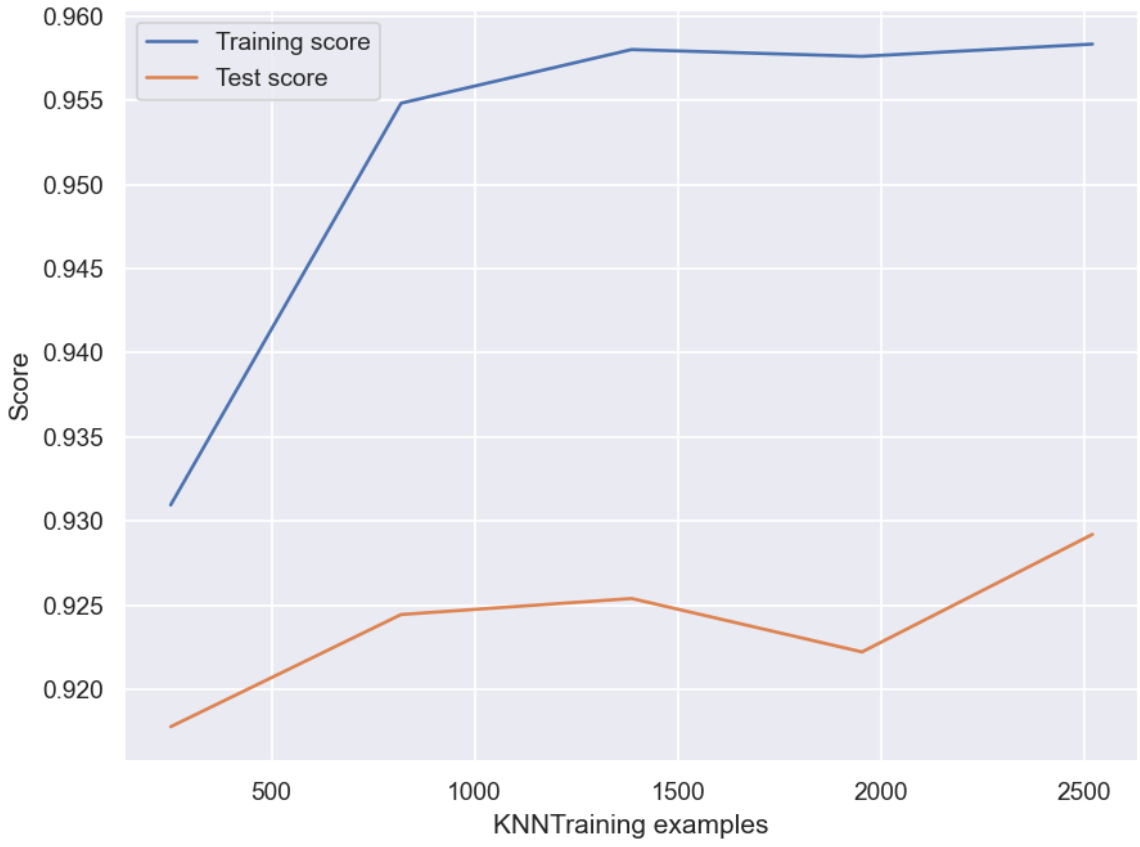
In this section, the Classification Algorithms will be used in the Customer Churn dataset to predict Churn attribute. At the end of this section a comparison of performance for the three models will be presented.

K-Nearest Neighbors

The below table shows the confusion matrix of applying K-nearest neighbor's classifier into the customer Churn dataset to predict Churn:

	Predicted Positive	***Predicted negative***
Actual positive	*52*	*52*
Actual Negative	*5*	*521*

The k-NN classifier achieved a precision of 0.91 and the ROC/AUC score was 0.96. The performance of the k-NN classifier on the customer churn dataset was relatively good, with a high precision score and a relatively high ROC/AUC score.



In the above graph, the difference between the test error and the training error is shown. KNN has caused high variance. Means it has an overfitting problem.

Naive Bayes

The below table shows the confusion matrix of applying Naïve Bayes classifier into the customer Churn dataset to predict Churn:

	Predicted Positive	***Predicted negative***
Actual positive	*96*	*8*
Actual Negative	*164*	*362*

The Naïve Bayes classifier achieved a precision of 0.98 and the ROC/AUC score was 0.90. The performance of the Naive Bayes classifier on the customer churn dataset was excellent, with a very high precision score and a high ROC/AUC score.

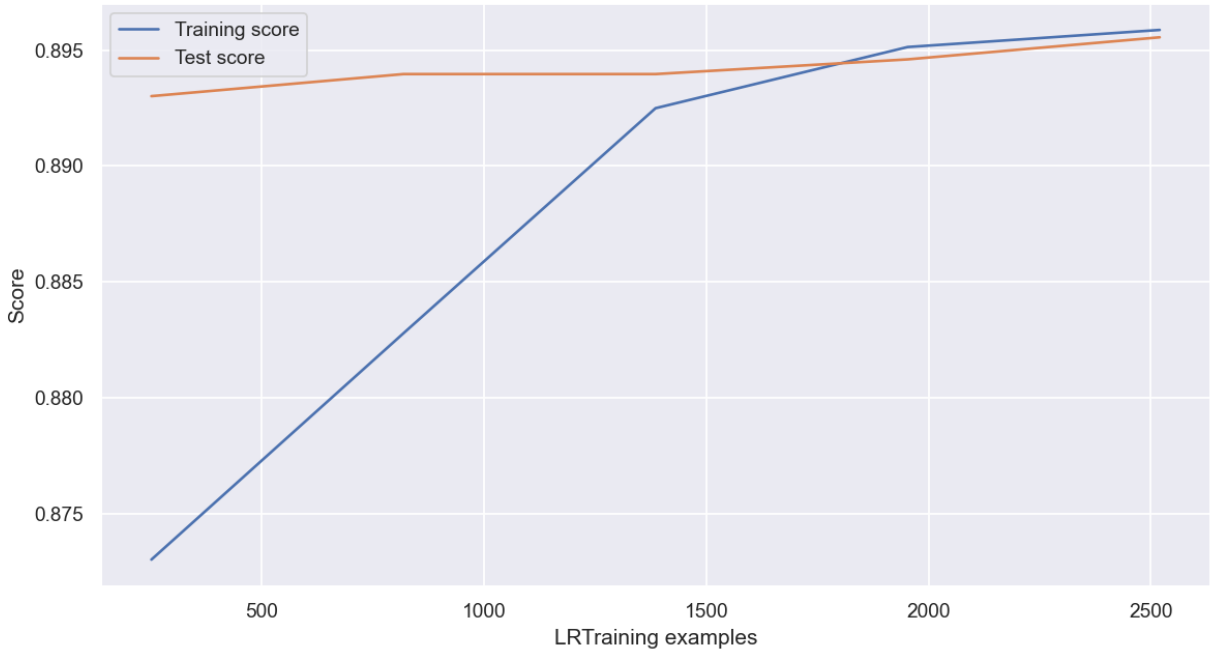
The training score and the test score keeps changing up and down when the size of the data is getting larger. Until they reach at some point steady changing, but the training score was higher than the test score. This might indicate under fitting or failing to follow the pattern.

Logistic Regression

The below table shows the confusion matrix of applying Logistic regression into the customer Churn dataset to predict Churn:

	Predicted Positive	***Predicted negative***
Actual positive	*38*	*66*
Actual Negative	*5*	*517*

The Logistic regression classifier achieved a precision of 0.88 and the ROC/AUC score was 0.92. The performance of the Logistic Regression classifier on the customer churn dataset was very good, with a relatively high precision score and a high ROC/AUC score.



In the above graph, the difference between the test error and the training error is shown. LR has caused high bias. Indicating under fitting.

Comparison of Performance

Below is a table that compares the performance of the Logistic Regression, Naive Bayes, and k-NN classifiers on the given dataset using ROC/AUC and precision:

	k-KK	***Naïve Bayes***	***Logistic Regression***
ROC/AUC	*0.96*	*0.90*	*0.92*
Precision	*0.91*	*0.98*	*0.88*

Based on the table it is noticed that the KNN classifier had the highest ROC/AUC score and NV scored the highest precision among the other classifiers. Then, the Logistic Regression classifier was the second-highest ROC/AUC score and scored precision 0.88.

But considering the learning graph results, the KNN classifier was an overfitting model.

Conclusion

In conclusion, the analysis of the customer churn dataset has shown that ID and Age group attributes will not contribute to the prediction of Customer value and the classification of Churn attribute so they were dropped.

It is also shown, that attributes selection, in which to decide the important factors to predict "customer value" will be more accurate if chosen based on the correlation matrix, and not on the point of view of the analyzer. Which were these attributes:

1- *Frequency of SMS*
2- *Frequency of use*
3- *Seconds of use*

As for the classification models, the Naive Bayes classifier was found to be the best performer among the three classification algorithms tested, because it is not an overfitting neither an under fitting model.

The results of this analysis can be used by your company to reduce customer churn. The consideration may be that the company starts using the Naive Bayes classifier as a tool to predict customer churn, hence prevent it. Also focus on improving the above attributes to make higher customer values.