

COMP4388

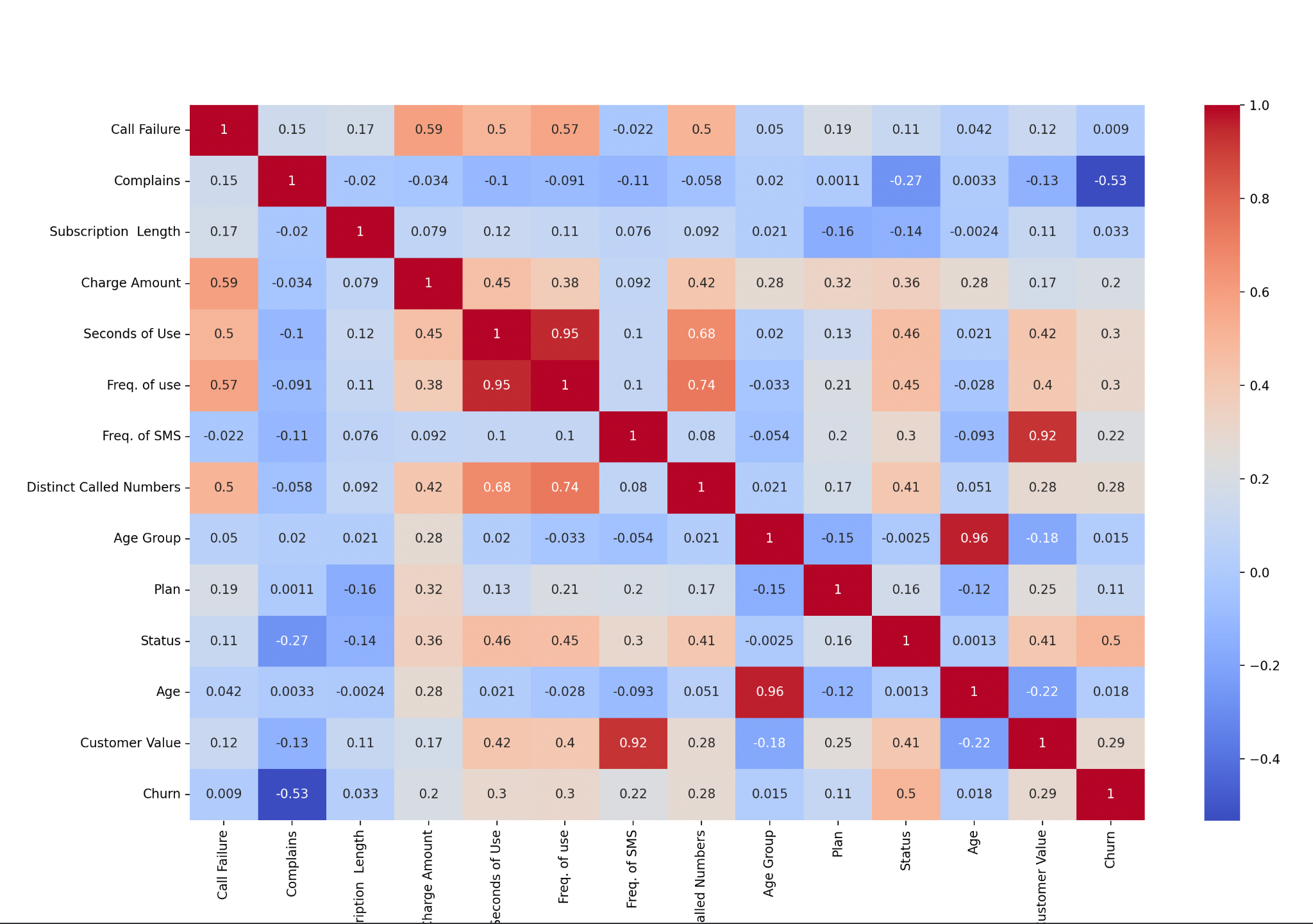
Project I Report

Dr Radi Jarrar

Maiia Kalajiawi (1170251)

Submitted onJan 2, 2023

# **EDA Task Analysis**



This heatmap was used to visualize the collaborations between all features. The color of each cell in the heatmap represents the strength of the correlation, with darker colors indicating stronger correlations and lighter colors indicating weaker correlations.

Two features that are strongly related are seconds of use and frequency of use with a relationship of 95%. Another two features that are strongly correlated is between age and age group with also a relationship of 95%. The last pair of strongly correlated features is customer value frequency of SMS albeit with a relationship of 92%. In addition, this strongly implies that changes in one feature tend to be accompanied by similar changes in the other feature.

The weakest correlation is between Churn and Complains with a relationship of -53%. Therefore we can conclude that they are independent from each other and do not affect each other.

# **Regression Task Analysis**

In task 2, I have chosen the three most important features in the dataset are: Call Failure, Status, and Freq. of use.

The justification behind the first is that while the customer might have the option to write a complaint, the customer most likely would not do so due to being reserved, busy, or non-communicative. This is why I think that instead of looking at the number of complaints, looking at the number of times where a service has failed would give us better insight towards a customer’s discomfort hence a likely cause to turn away from this service.

As for the second feature, I have chosen Status to factor in the current activity status of the customer that could hint towards an early decision to discontinue the service. If the customer is not currently using this service, it could be very likely he intends to discontinue it.

Regarding the feature of Freq. of use, the customer might have some thoughts on the product’s experience the more frequently they use it. A particular feature might be lacking to a user that will eventually move away from this service once it’s found in another service.

| **Model** | **R2** | **MSE** | **MRSE** | **MAE** | **MDAE** | **MDAEP** |
| --- | --- | --- | --- | --- | --- | --- |
| LRM1 | 0.98 | 1.83 | 1.35 | 0.78 | 0.49 | 0.16 |
| LRM2 | 0.88 | 11.96 | 3.4 | 2.38 | 1.55 | 0.62 |
| LRM3 | 0.96 | 3.78 | 1.94 | 1.04 | 0.34 | 0.21 |

From this table, we can conclude that LRM1 has the least MAE value and also the highest R2 making it the best model.

Mean squared error (MSE): This metric measures the average squared difference between the predicted and actual values. A lower MSE indicates a better fit.

R-squared: This metric measures the proportion of the variance in the target variable that is explained by the model. A higher R-squared indicates a better fit.

Mean absolute error (MAE): This metric measures the average absolute difference between the predicted and actual values. A lower MAE indicates a better fit.

# **Classification Task Analysis**

| **Model** | **Accuracy** | **ROC/AUC** |
| --- | --- | --- |
| KNN | 0.93 | 0.87 |
| Naive Bayes | 0.73 | 0.81 |
| Logistic Regression | 0.90 | 0.71 |

From this table, we can conclude that KNN has the best accuracy and is well-suited for this type of data.

# 

# **References**

1. Python. (2023). *Python Documentation 3.11.2*. https://docs.python.org/3/