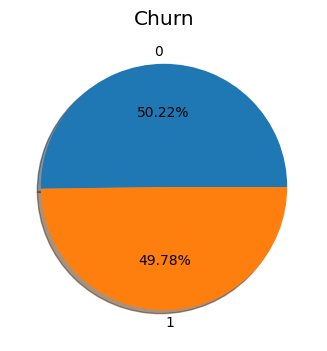
Customer Churn Prediction

Sreevaatsav.B

# Data preprocessing:-

Firstly, the dataset had equal distribution of samples in both the classes.

The data had no NULL values when checked and it was split into training and testing parts with resting size approximately 15%.

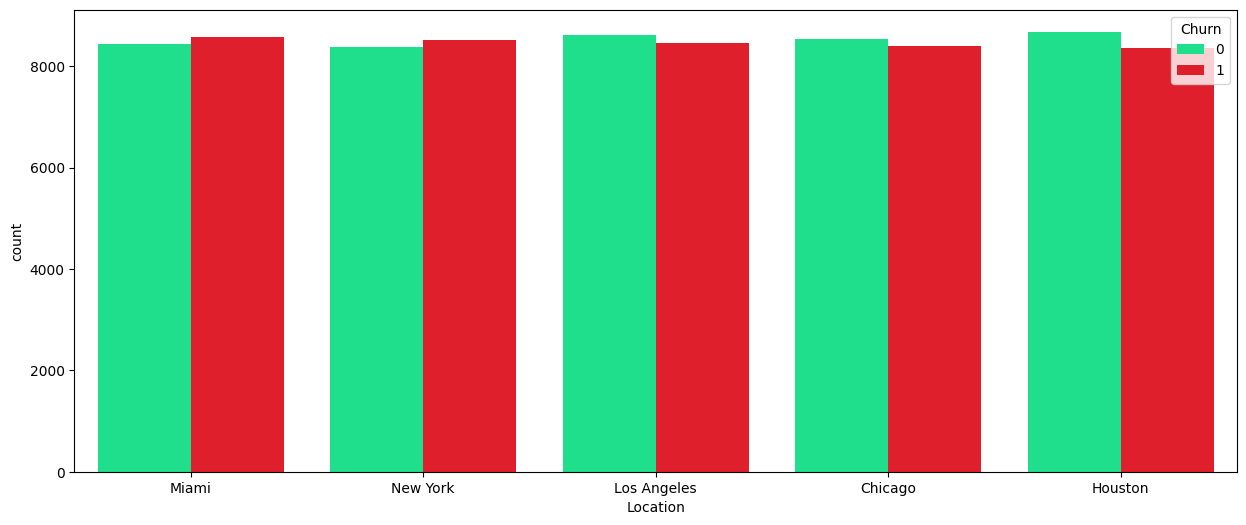
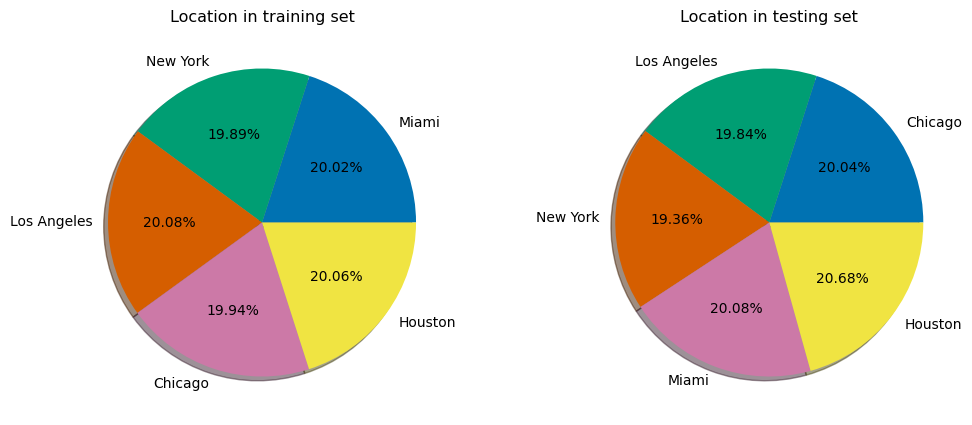
The categorical features like Gender and Location were encoded into integers by assigning each of the values a unique numeric value.

The features like name and customer id is not used to model the data, as they won’t affect the customer’s decision to leave the service.

# Exploratory Data Analysis:-

**Location:-**

The location of customers are mainly spread uniformly in 5 different places as shown in the below graph.

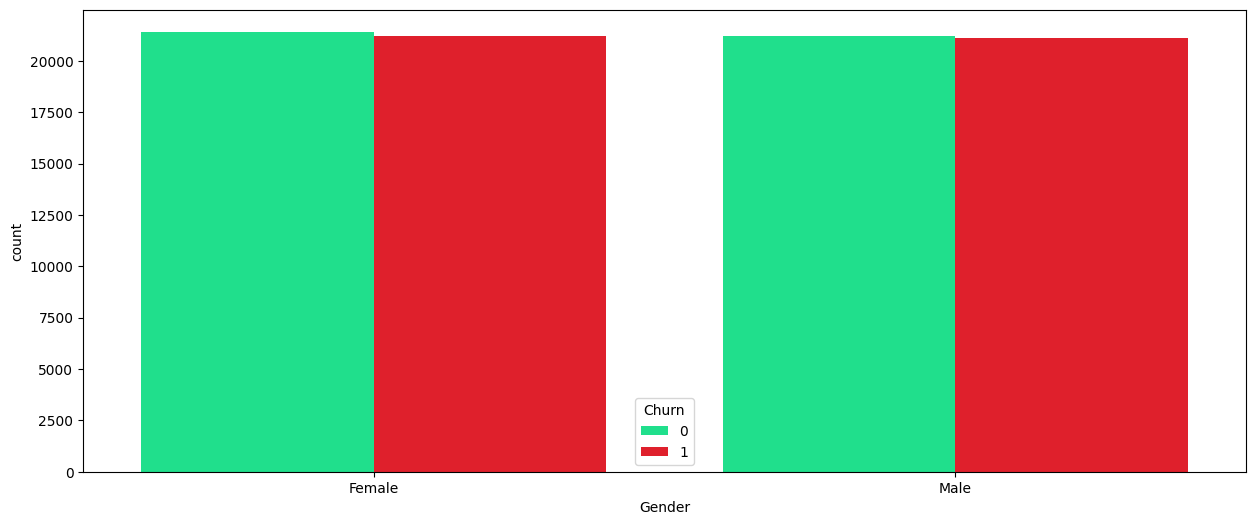


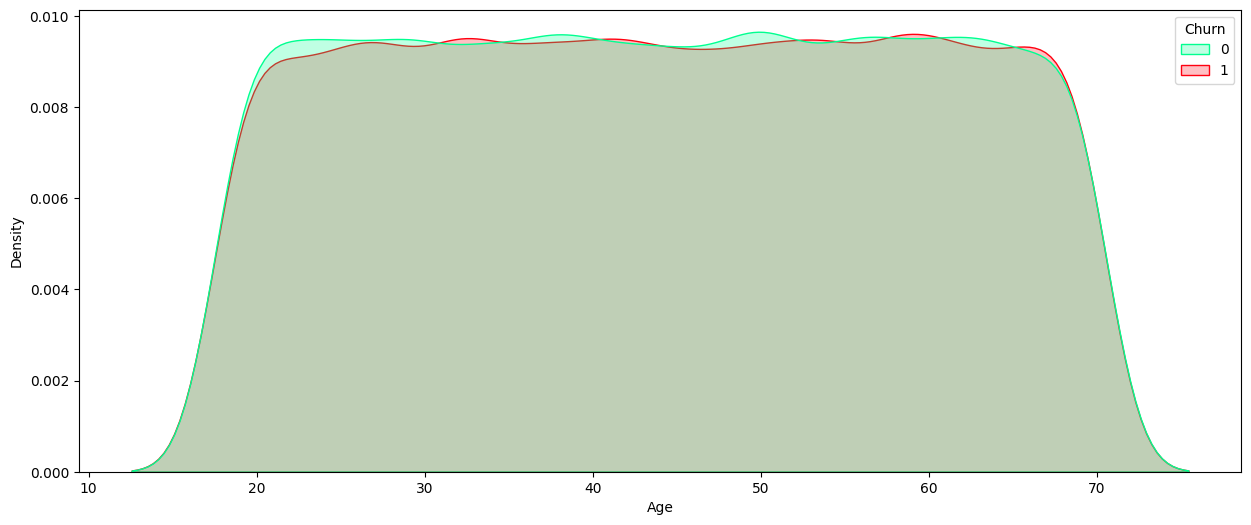
As seen from the above figure, the distribution of classes is shown based on the location of a user, and we can see that there is almost equal distribution of the classes and not a lot of difference can be seen.

**Gender :-**

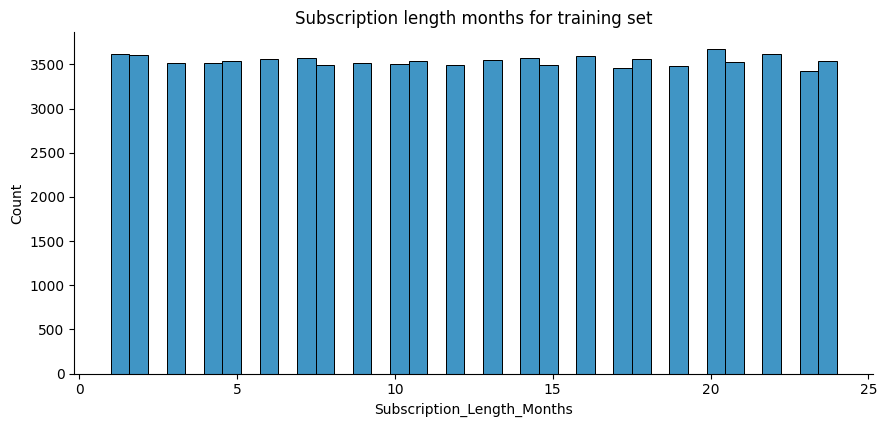
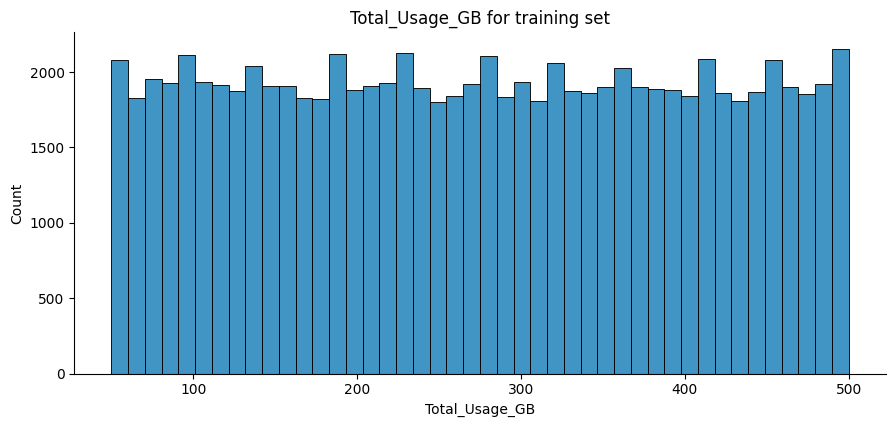
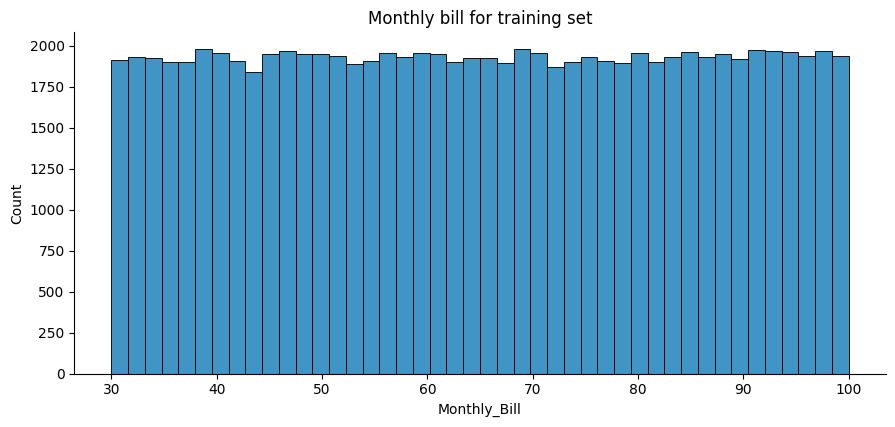
The genders distribution in both the classes is almost same as shown in the figure below.

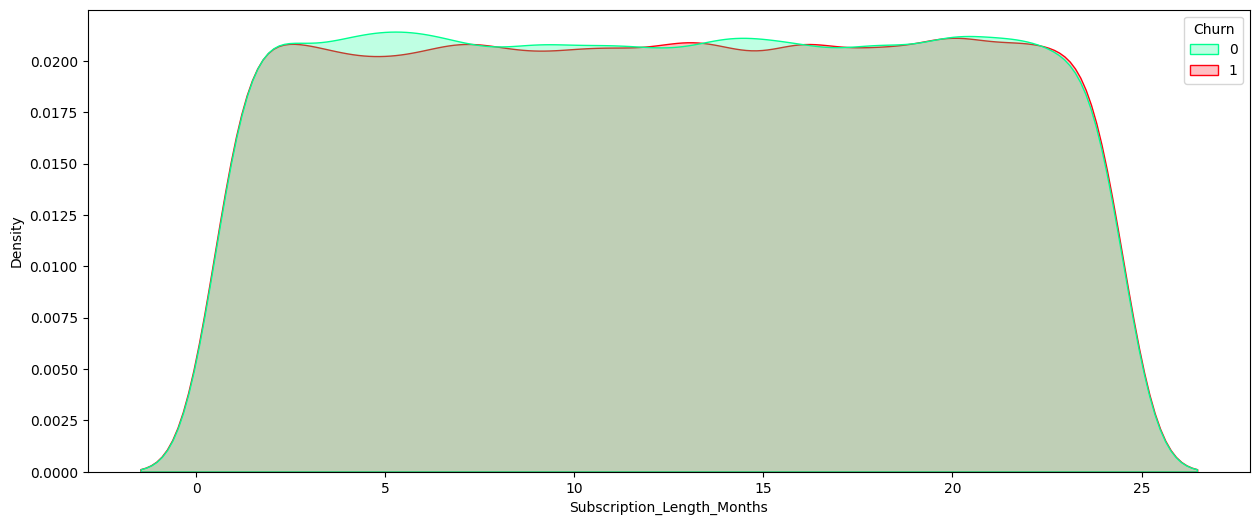
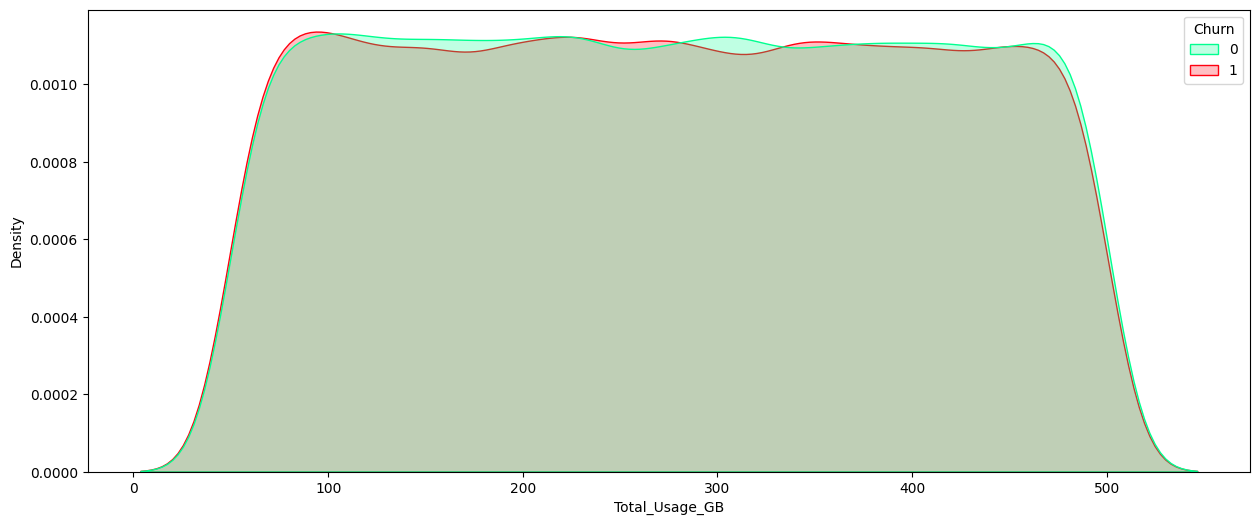
**Age:-**



As the previous features, even the age of the customers doesn’t vary with the class, and can be seen form the density plot shown below, they almost overlap.

**Subscription length, Total data usage, Monthly bills:-**

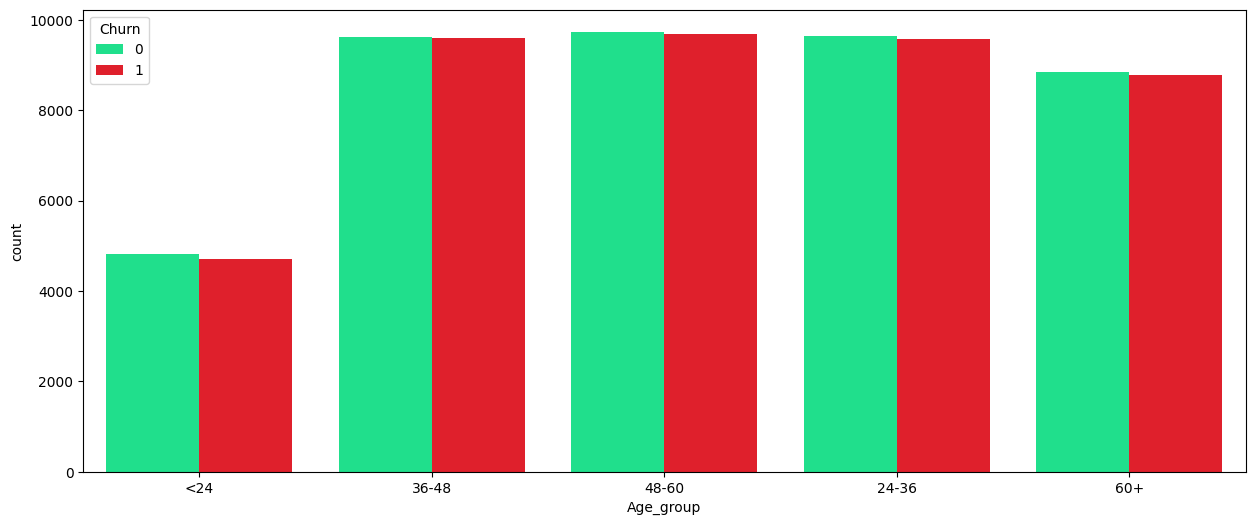
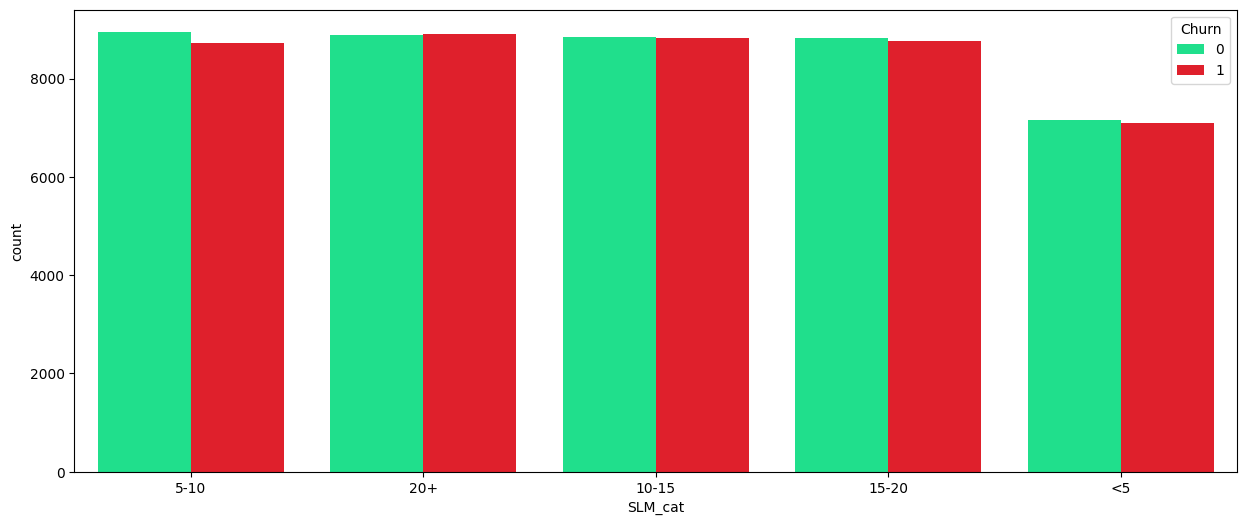
All these features are distributed uniformly in their respective range of values and it’s the same as the above features, where there is no significant difference between the distributions of the 2 classes at all.

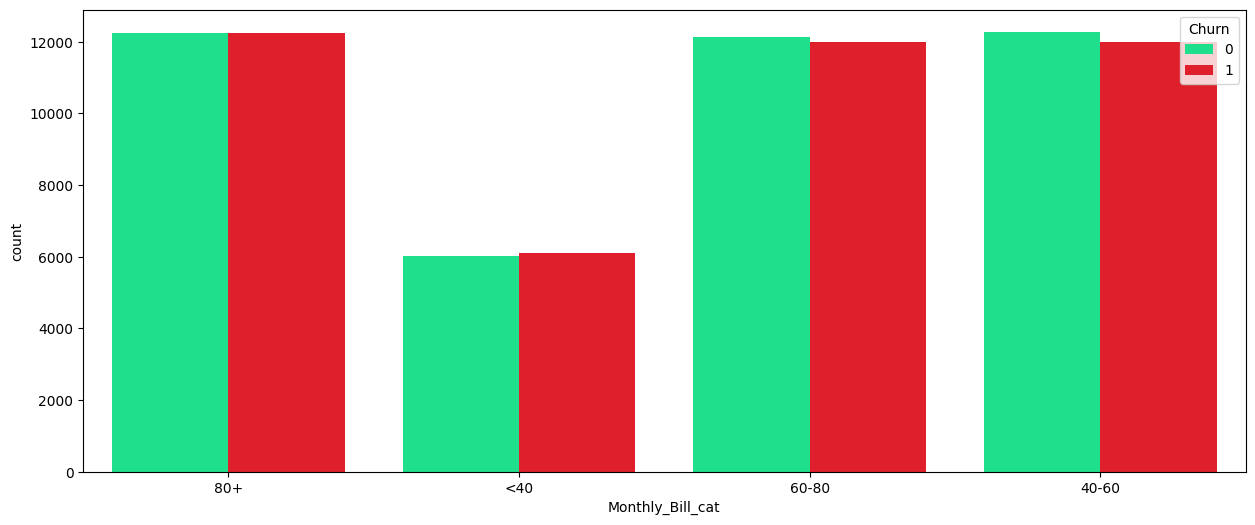
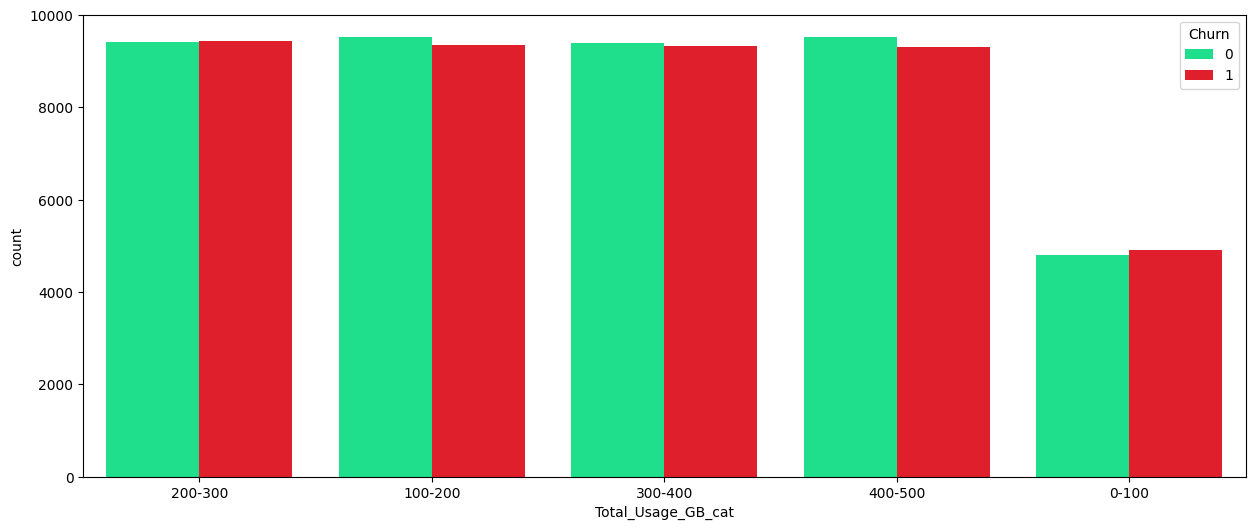
The figures are shown above respectively.

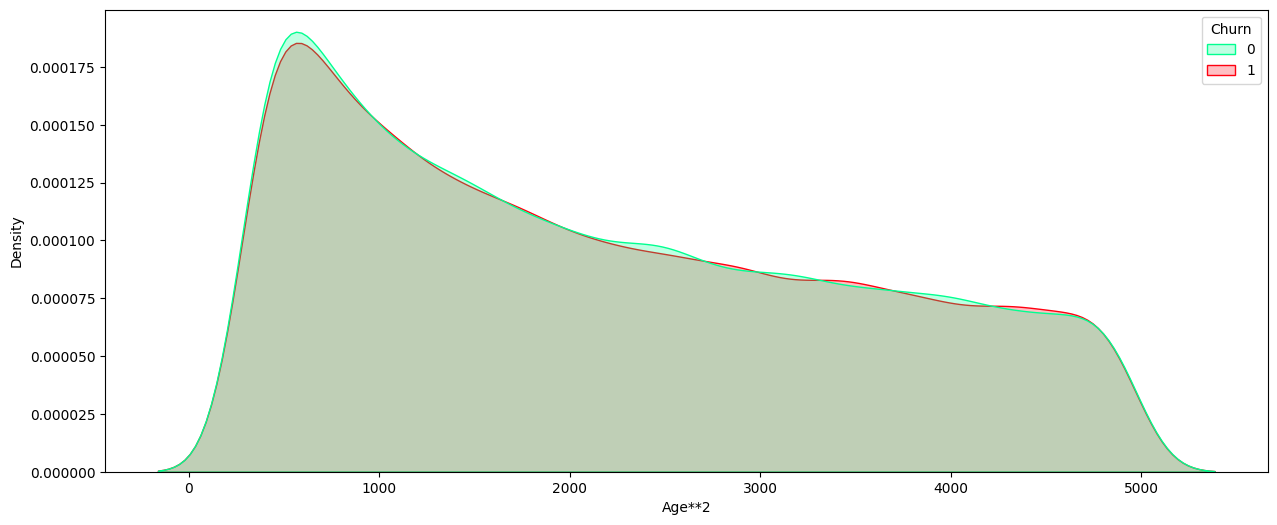
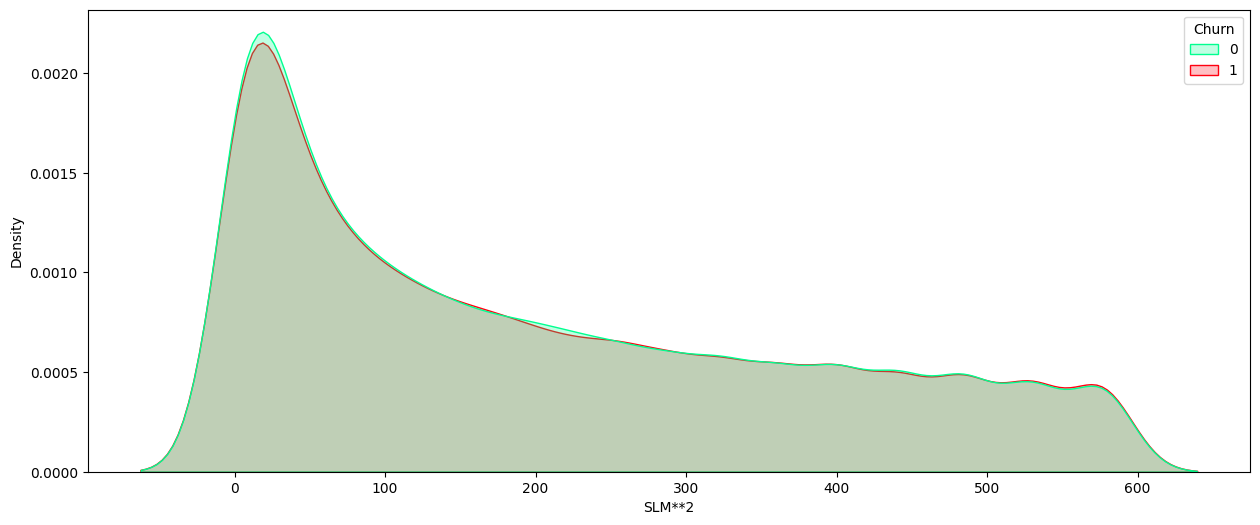
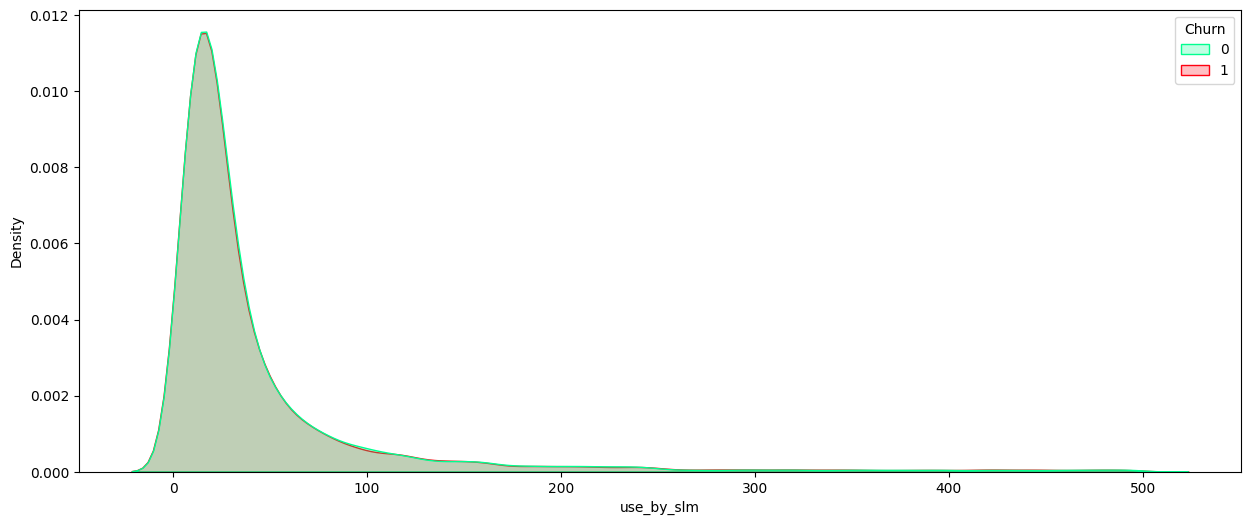
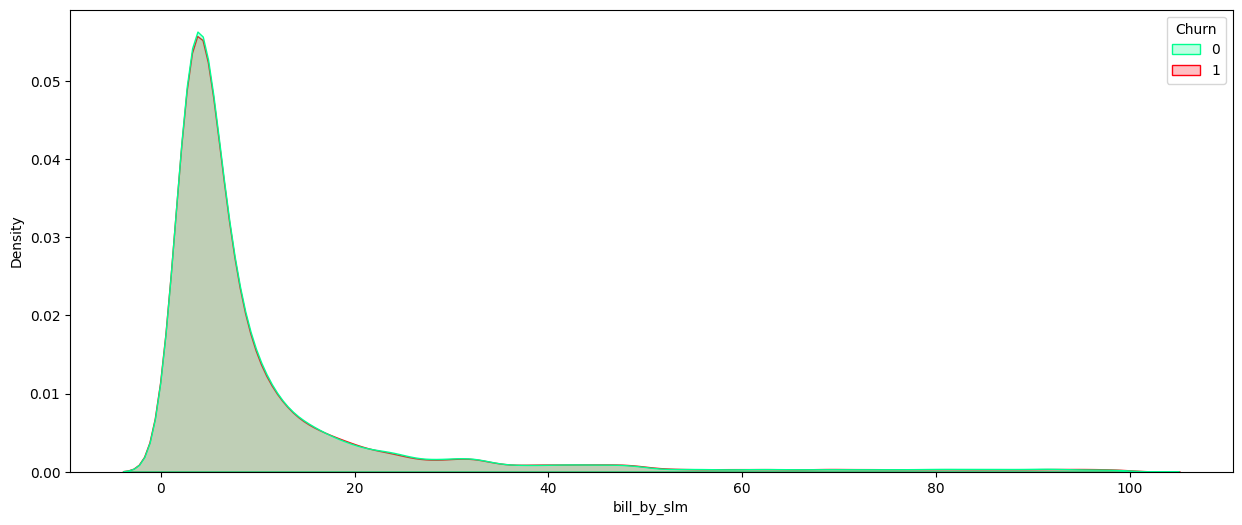
Due to these kind of overlap of features between the 2 classes, I did some feature engineering on the categorical and numeric features independently.

# Feature engineering:- monthbills_cls.png

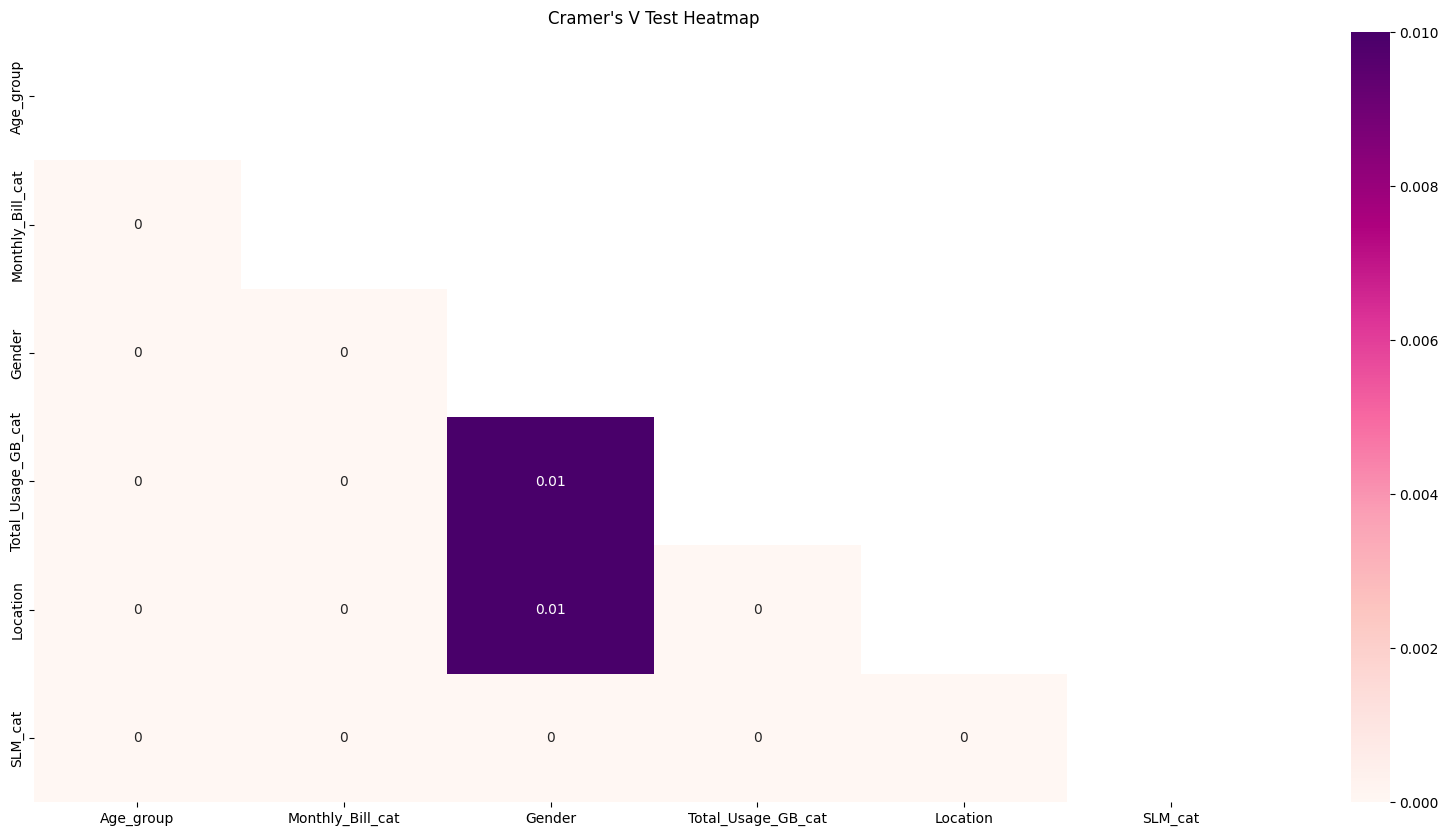
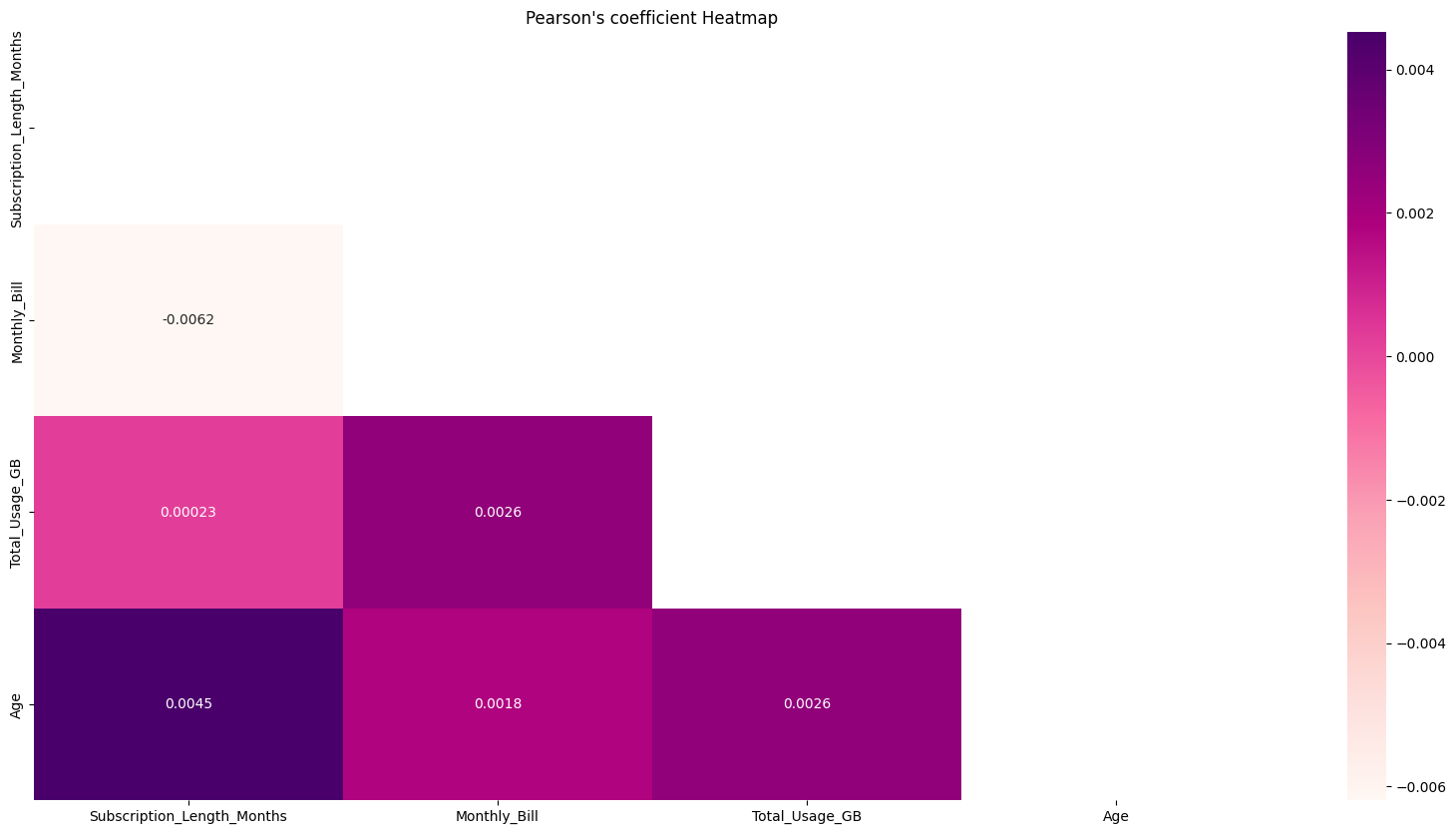
**Numerical features engineering:-**

As the distributions are uniform, I have split the range of feature values in to several splits and can be seen in the figures below. Thus the numeric features will be categorised and be treated as categorical features.



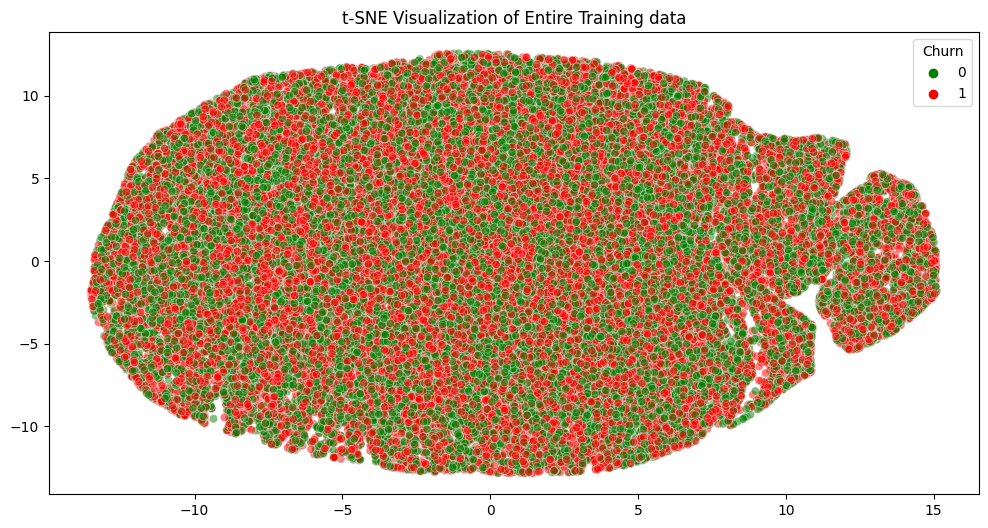
While for some other features like, Age and subscription length were squared and taken as new features, while new features were added like (data usage)/(subscription length), (monthly bill)/(subscription length). The figures are shown above. This is done after observing the performance of the models trained on only the categorical features.

The Pearson’s correlation between the original numeric features:-

The Cramer’s V test for the categorical features (for constructed features as well):-

From the above plots, we can see that magnitudes of any of the feature correlations are near 0, suggesting independence between features. This is mostly due to the uniform distribution of data among all the features independent of the class.

**Data visualisation in 2D of the constructed numeric data:-**



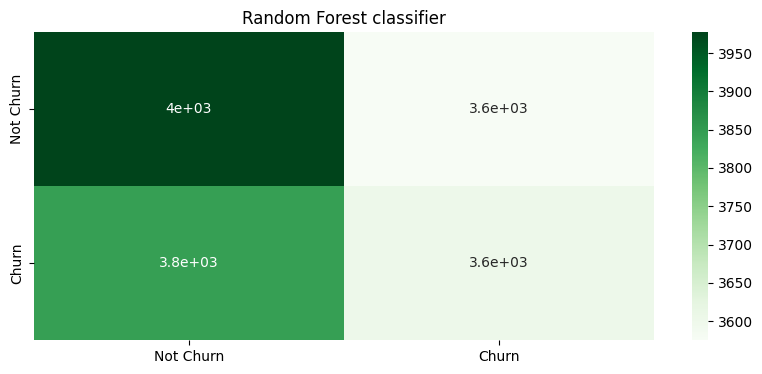
Here we can clearly see that the data point in 2 classes have no clear distinction or grouping to classify them. Thus we can expect only lower accuracies on the given data!

# Modelling:-

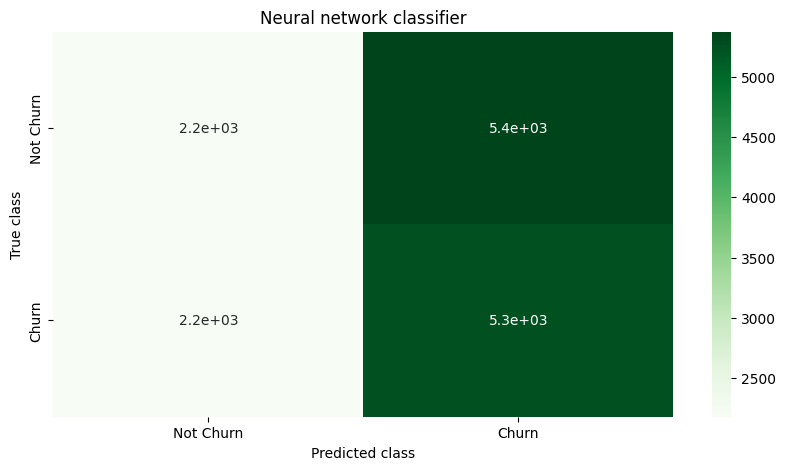
Firstly, only the categorical data (original + numeric converted to categorical) is used to model the data.

With algorithms which are based on trees, such as decision trees, random forests, light GBM and XGBoost are used to train and evaluate the data.

The best hyperparmeters for each of these models are chosen by considering a grid-search of cross-validation = 5.

With all these algorithms on this data, the maximum accuracy obtained is about 51%, with f-1 score of about 49%. (Randomforest with max\_depth = 2 and n\_estimators = 10)

As we can see, these algorithms did not perform so well on the churn class, so only the numeric features are then used for training a new model.

By experimenting with several model such as logistic regression, tree-based algorithms, a neural network with 1 hidden layer with ReLU activation functions, performed in similar way with much better recall for the churn class. (71% recall on churn class)

Even this model isn’t great in terms of classifying the data points.

Another neural network is trained only by using the all the original numeric and categorical features (without any new features). As a recall of about 0.6 for the churn class with accuracy 0.5.

**These 3 models are used as an ensemble with majority class as the final prediction.**

The blend of these models has a recall of 0.63 for the churn class and accuracy of 0.50.

This accuracy is due to the type of data that was given.