Final Report

for

Customer Churn Analysis

Version 1.0

Prepared by

Ayesh Vininda

Contents

1. Problem Overview

Customers are the foundation of any business success. That is why every company tries to do anything to make their customers satisfaction. Customer churn is loss of customers from the business, and it is not good for the endurance of the company in the competitive business ground. Organizations must devise several solutions to address churn issues based on the services they provide. Chatter Box is a Telco company, means Telecommunication or communication service providing company.

Customer churn management is critical in the competitive and continuously evolving telecom industry. Customers are rapidly change their telco service providers due issues of the services, service charges and other benefits and rewards giving from other companies. Even customers are churned new customers are quickly joined. But the major problem is also acquiring new customers. Because of acquiring new customer is more cost than retaining existing. So, acquiring, and immediate churning is very unpleasant situation for the telco companies. So, telco companies should find the way hold their customers.

In our case Chatter Box also has this problem. To solve this problem, we must use descriptive analytics, diagnostic analytics, and predictive analytics. Using those analytics techniques Chatter Box company hopes to reduce customer attrition rate and get the real advantages of acquiring new customers. Using these approaches Telco company hopes to get the view of the problem and predict about the customer churn or not.

To identifying the problem and the solution Chatter Box company hopes to use Data Science and Machine Learning approach throughout visualization of analytics and prediction model to estimate the possibilities to churn on the web-based application. The purpose of building this kind of analytics and predictive application is to make the easy on decision making to improve customer retention.

2. Dataset Description

Dataset Name	Customer churn train and test datasets.
Dataset Size	311KB
No.of variables	Train dataset – 19, Test dataset - 18
No.of Data entries	Train dataset – 2321, Test dataset - 1500

Variable Name	Description	Data type
customer_id	Customer identification number	Categorical - Nominal
account_length	Number of months the customer has been with the current telco provider	Metric - Discrete
location_code	Customer location code	Categorical – Nominal
international_plan	If the customer has international plan or not	Categorical - Nominal
voice_mail_plan	If the customer has voice mail plan or not	Categorical - Nominal
number_vm_messages	Number of voice-mail messages	Metric - Discrete
total_day_min	Total minutes of day calls	Metric – Continues
total_day_calls	Total number of day calls	Metric - Discrete
total_day_charge	Total charge of day calls	Metric - Continues
total_eve_min	Total minutes of evening calls	Metric - Continues
total_eve_calls	Total number of evening calls	Metric - Discrete
total_eve_charge	Total charge of evening calls	Metric - Continues
total_night_minutes	Total minutes of night calls	Metric - Continues
total_night_calls	Total number of night calls	Metric - Discrete
total_night_charge	Total charge of night calls	Metric - Continues
total_intl_minutes	Total minutes of international calls	Metric - Continues
total_intl_calls	Total number of international calls	Metric - Discrete
total_intl_charge	Total charge of international calls	Metric - Continues
customer_service_calls	Number of calls to customer service	Metric - Discrete
Churn	If the customer left or not (target variable)	Categorical - Nominal

3. Data Pre-processing

The dataset has customer id, but it is not variable which affect to the target variable. So, we don't consider it as feature for our dataset.

3.1 Data cleaning

Remove Unwanted Columns

- Datasets show some columns such as 'Unnamed: 20','Unnamed: 19', which are not in data description. These are kind of errors when creating datasets. So, we can consider them as the unwanted columns and remove them.

Remove duplicates

- In the train dataset has some duplicate entries. If we train our model with duplicates, then model will overfit. So, we must remove those duplicates.

Handle Missing values, Outliers, Invalid data

- In both train and test dataset have missing values. It is hard to train model with missing values because of lot of models which are used in this project (machine learning models in sklearn) are not supported to deal with missing values. So, we must handle missing values. But at this moment it is hard to handle the missing values because of missing values techniques are dependent on outliers and missing data.
- These datasets have outliers and invalid data.
    - According to dataset description any of these data cannot be negative. But there are some negative values. So, we can mark these negative values as the missing values (NaN s).
    - There are some clearly visible outliers in these datasets. To remove outliers, we are normally use quantile bounds. But in here we cannot use quantile bounds because quantile bounds cuts possible bounds also. So, I created suggested boundary for every numerical variable going through the test and train datasets comparing box plots of features in both datasets. Now these boundaries are removing the invalid and outliers from the dataset without removing possible cases.
    - Using suggested boundary set values as NaNs which values go over the boundary.
- Now mark all outliers and invalids as missing values. So, we can replace missing values by using some methods.
- Correlation matrix before the filling missing values shows high correlations sets So, we can use Regression Model for impute those sets.
- Other missing values of numerical variable imputed by median. And, there are few missing values in categorical variables we can impute them by mode.

3.2 Feature Engineering

Encoding Categorical data

- To train machine learning model we need numerical valued things. But there are some variables with not valued as numerical.
- So, we are normally using encoding techniques convert them into numerical valued things.
- voice_mail_plan, intetiol_plan and churn were encoded by Label encoding.
- location_code was encoded by one hot encoder.

Create new features

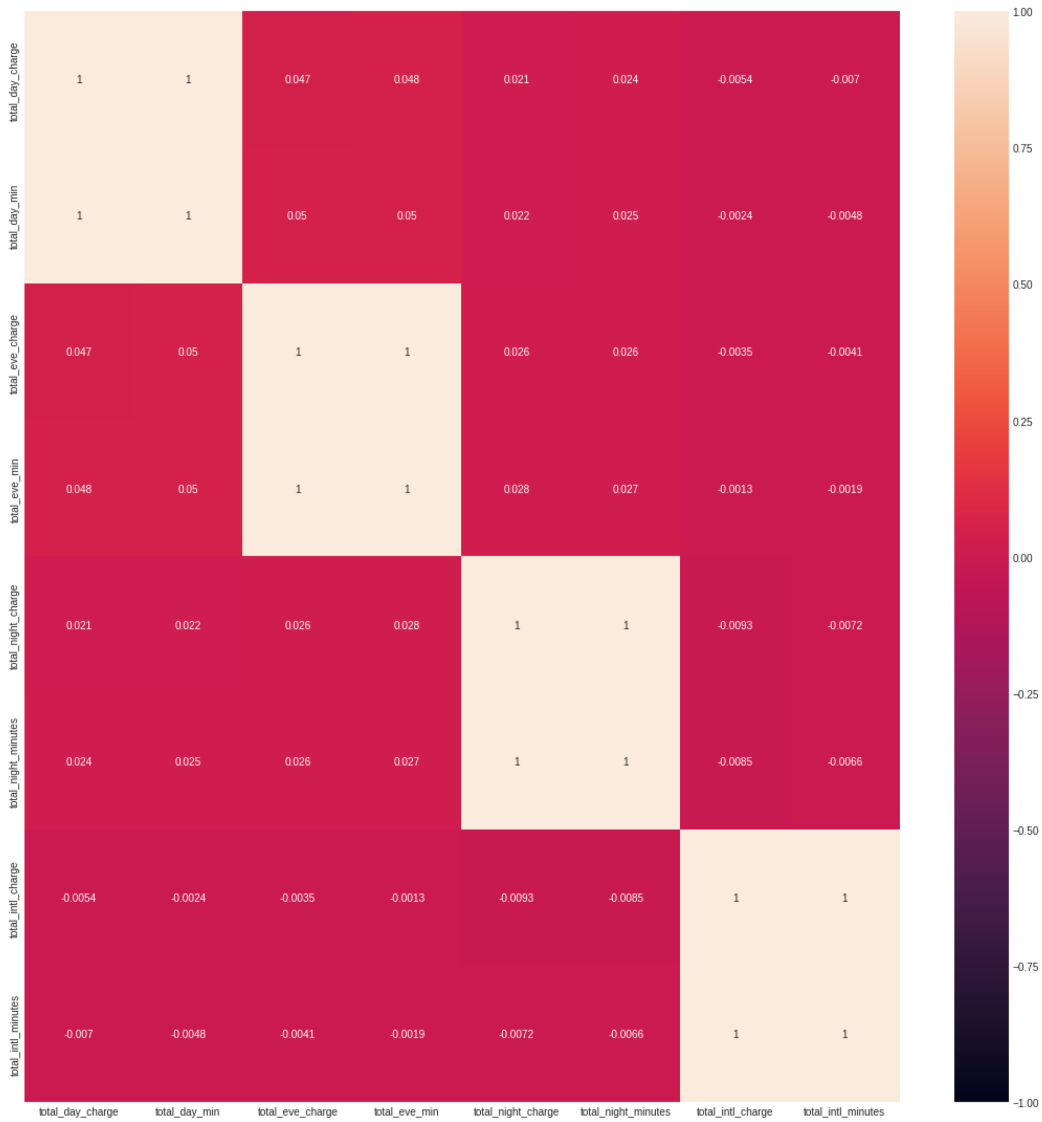
such as,

- no_of_plans = get the number of plans
- total_charge = total_intl_charge + total_night_charge + total_eve_charge + total_day_charge
- total_calls = total_intl_calls + total_night_calls + total_eve_calls + total_day_calls
- total_min = total_intl_minitues + total_night_minutes + total_eve_minitues + total_day_min

4. Insights from data analysis

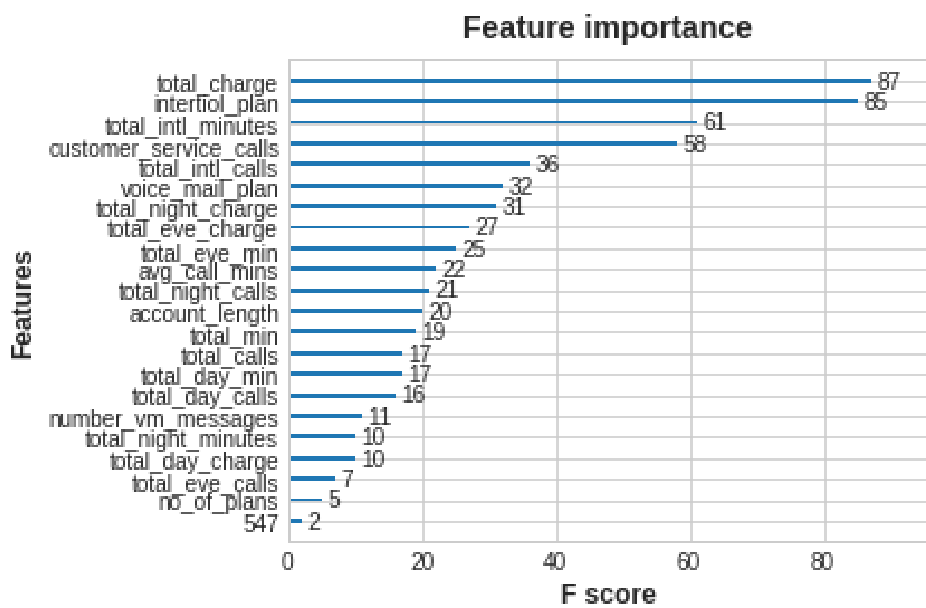
Insight 1

- As the first derived insight we can show the high correlation between some features.
- This is derived at after setting all outliers and invalid data as NaNs.
- So, one to one correlation can be used to compute NaN values.

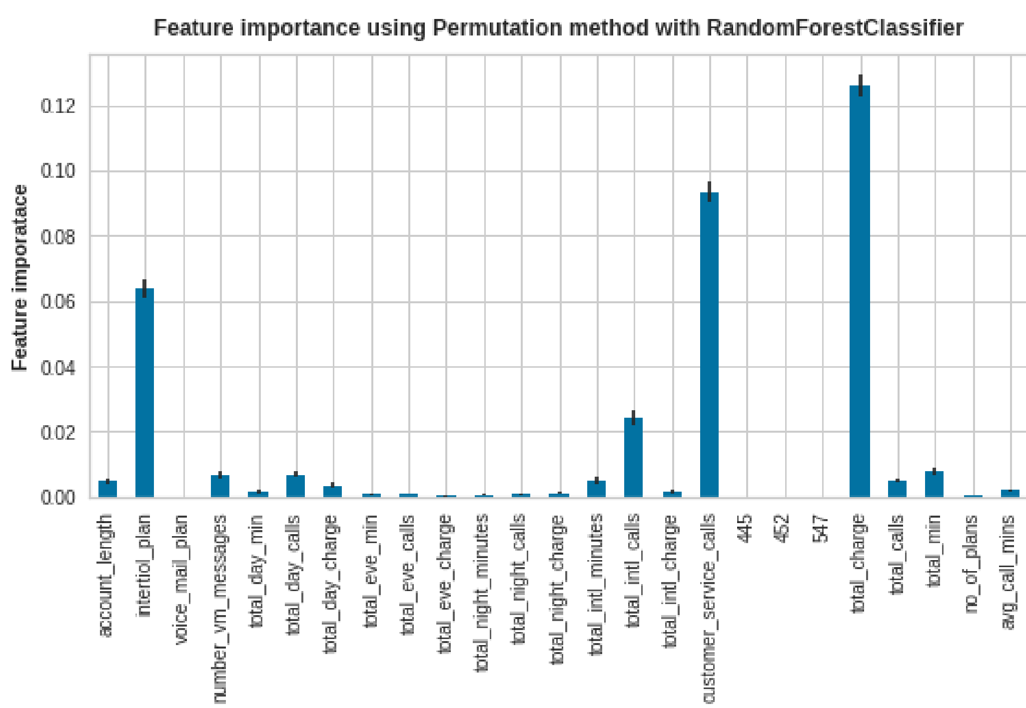


Insight 2

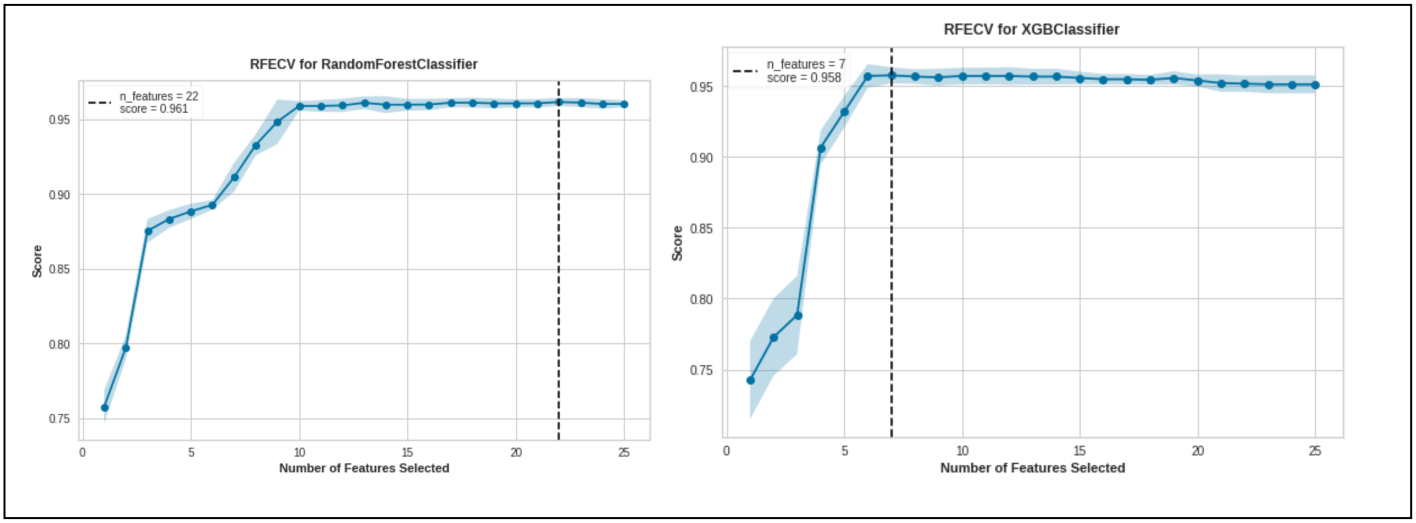
Identify most influential columns that increase Churn



Feature importance using XGBClassifier



- by looking at features importance generated by using XGBClassification and permutation with RandomForestClassifier model we select total_charge, intretiol_plan, total_intl_minitues , customer_service_calls, total_intl_calls five most influential features that increase customer churn.

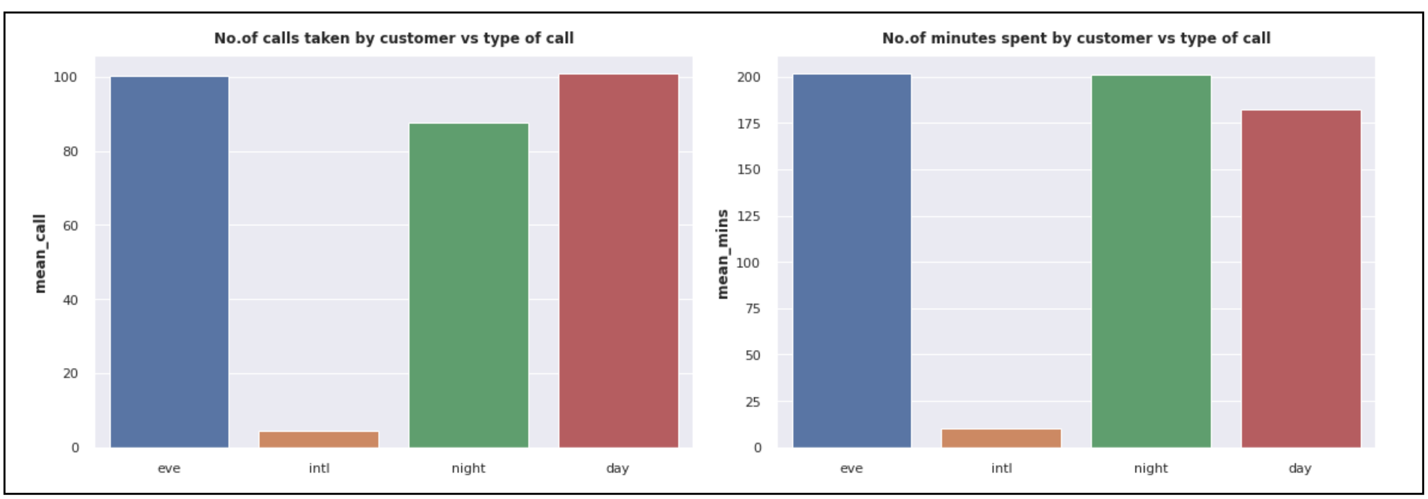


- Using recursive feature importance for RandomForesetClassifier and XGBClassifier can used to select how many numbers of features to use.

Insight 3

Type of call customer prefer to take.

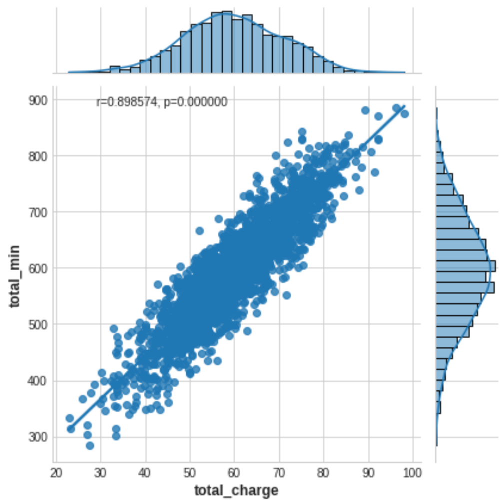
- The dataset has 4 types of calls.
  - Evening calls
  - Night calls
  - Day time calls
  - International calls
- So, which type of calls customers most prefer to get.



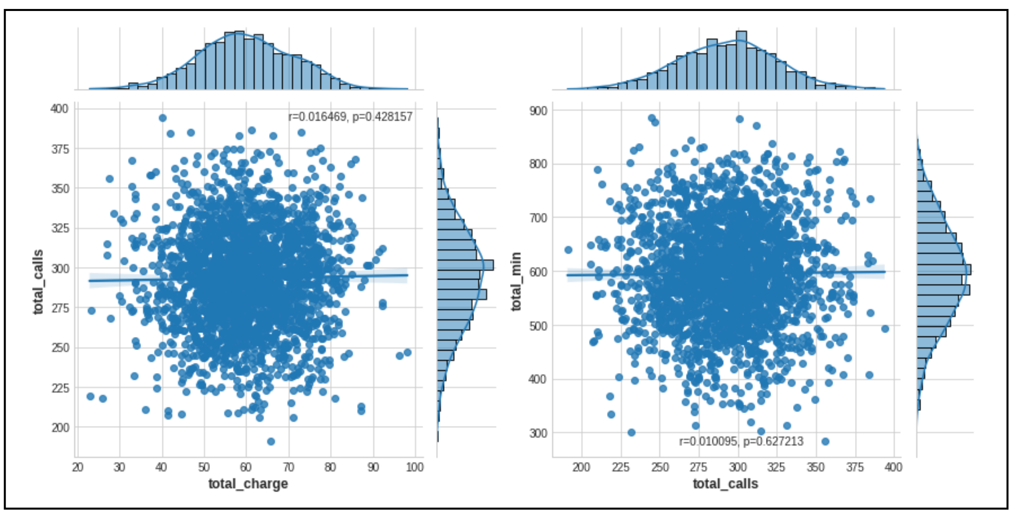
- ❖ by looking at the overall customer graph a smaller number of customers are taken international calls. And most of customers prefer to take day and evening calls. And they are spent more time at evening and night on calls.

Insight 4

Correlations between total_min, total_charge, total_calls only total_min and total_charge got higher Pearson correlation.



❖ According to Pearson correlation there is higher correlation between total mins.
❖ Positive Pearson correlation value is 0.8985.
❖ So, between total_min and total_charge has good linear relation.



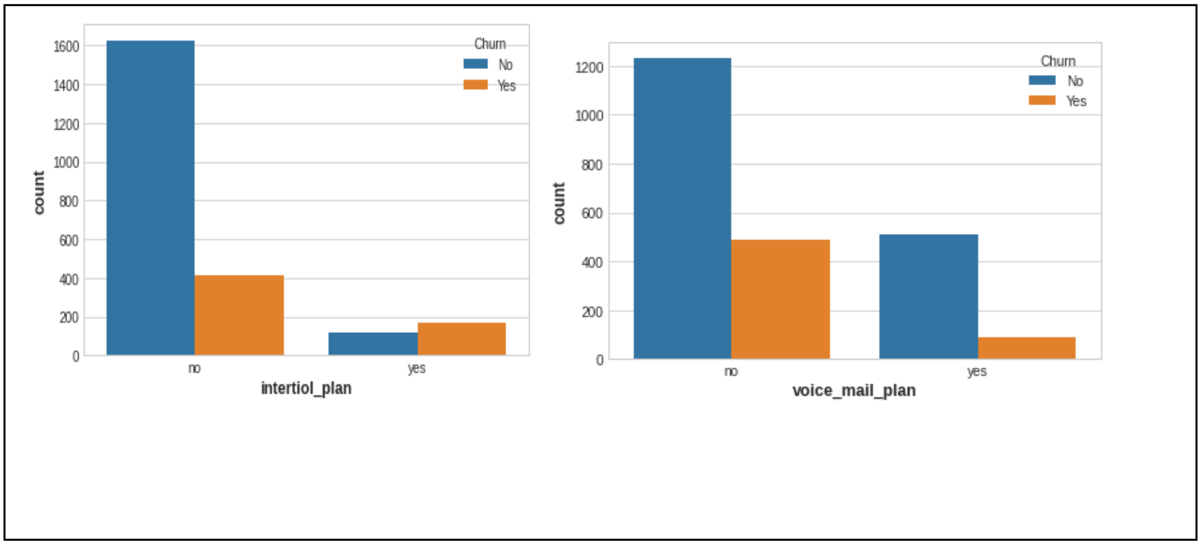
- ❖ Pearson correlation of
  - o total_charge vs total_calls :0.01646
  - o total_min vs total_calls : 0.01009

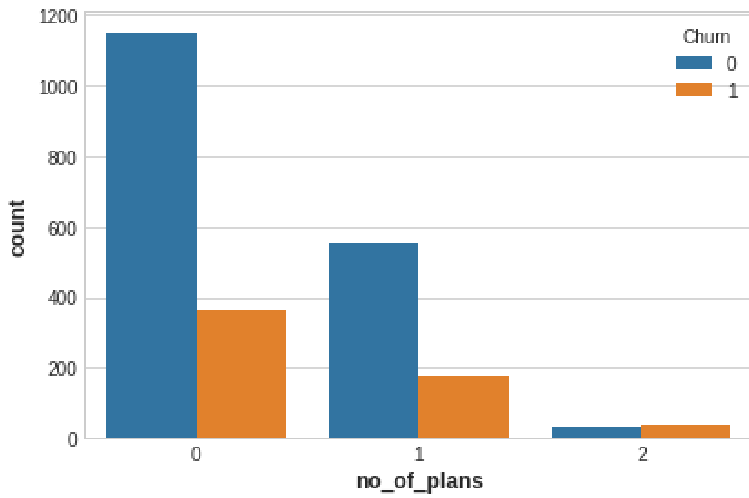
    are very low values.

    So, no linear relationships between them.

Insight 5

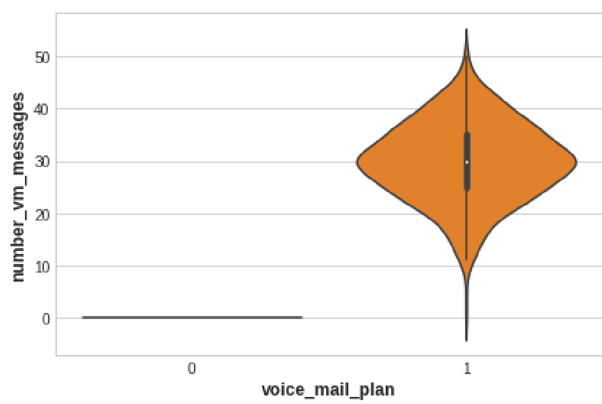
Most of customers who selects voice_mail_plan prefer to stay with the service.





- Most of customers prefers to without having any plan.
- Customer who has one plan prefer to stay with the service.
- Most of customers who has international plan prefer to leave the service.
- Most of customers who are select a plan, stay with the service because of they select voice_mail_plan as their plan.

==



Customers who don't have voice mail plan don't send any voice massages.

==

- In location code 452 larger number of people are using this service.
- And also, location code 452 shows large number of churns.