Project report on

Customer Retention Retail

# Data set and Domain

## Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Definition** |
| custid | object | Computer generated ID to identify customers throughout the database |
| retained | int64 | 1, if customer is assumed to be Retained , 0 = Not Retained |
| created | datetime64 | Date when the contact was created in the database - when the customer joined |
| firstorder | object | Date when the customer placed first order |
| lastorder | object | Date when the customer placed last order |
| esent | int64 | Number of emails sent |
| eopenrate | float64 | Number of emails opened divided by number of emails sent |
| eclickrate | float64 | Number of emails clicked divided by number of emails sent |
| avgorder | float64 | Average order size for the customer |
| ordfreq | float64 | Number of orders divided by customer tenure |
| paperless | int64 | 1 if customer subscribed for paperless communication (only online) |
| refill | int64 | 1 if customer subscribed for automatic refill |
| doorstep | int64 | 1 if customer subscribed for doorstep delivery |
| favday | object | Customer's favourite delivery day |
| city | object | City where the customer resides in |

### Variable categorization (count of numeric and categorical)

Number of Numerical columns - 11

Numerical column names - retained, esent, eopenrate, eclickrate, avgorder, ordfreq

, paperless, refill, doorstep, create\_first, first\_last Number of Categorical columns - 2

Categorical column names - favday, city

**Pre-Processing Data Analysis (count of missing / null values, redundant columns, etc.)**

The dataset indicates the overall purchase history of the customers who have purchased from the Online Tea Retail Store. The Data set consists of different fields which might consist of different parameters which affect the customer churn rate.

The online tea retail store data considered here is from 2008 and 2018 which consist of around 30801 observations and has 15 columns. The online tea retail store sells tea of different flavors across various cities in India. The dataset contains data about the store's customers, their orders, quantity ordered, order frequency, city, etc.

The Dataset consists of some missing values. The Columns which have around 20 missing values include Custid, Created, firstorder and lastorder. The column Custid contains the serial number of the customer, which is redundant for further analysis. Thus, we drop the column. We have Dropped a few observations where the format was not appropriate (‘'1/0/00', '00:00:00'). Since the Null value count was very negligible as opposed the count of total observations we have dropped that as well.

The columns firstorder and lastorder had datatypes (Object) that were inappropriate and hence we have changed the datatype (Datetime64). We also observed that the data is skewed and not normal. Hence, we will be using Transformation techniques to treat that.

We also observed an imbalance in the data where the retained customers consisted of 79.46% and churned customers consisted of 20.54%. We have observed that few of the columns had outliers and we have used the IQR technique to impute the outliers using Boxplot.

## Problem Statement:

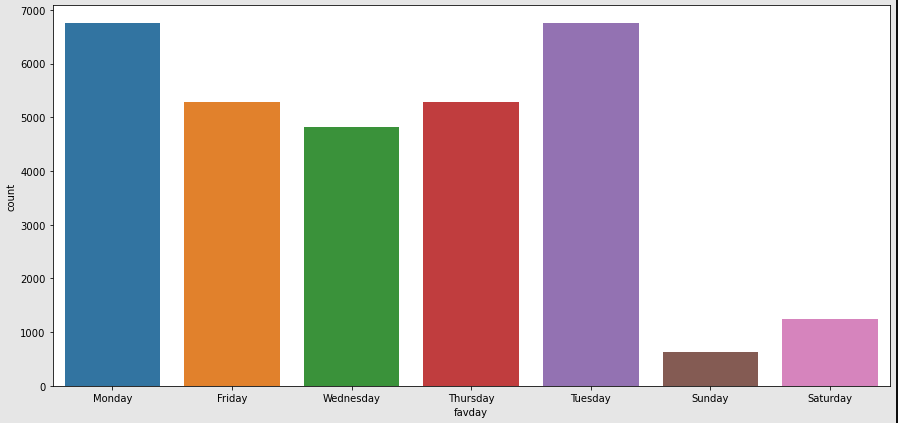
The Online Tea Retail store has encountered the problem of losing out existing customers in recent times. The Business owner isn’t aware of the reasons why the customers are churning and would like to bring down the churn rate. The churn rate, also known as the rate of attrition or customer churn, is the rate at which customers stop doing business with an entity.

### Target:

* To Increase customer retention by adopting different strategies.
* To come up with an effective targeted marketing strategy to avoid customer from churning.
* To formulate strategies for demographics of different locations.
* To understand customer behavior in deep manner.

# EXPLORATORY DATA ANALYSIS (EDA)

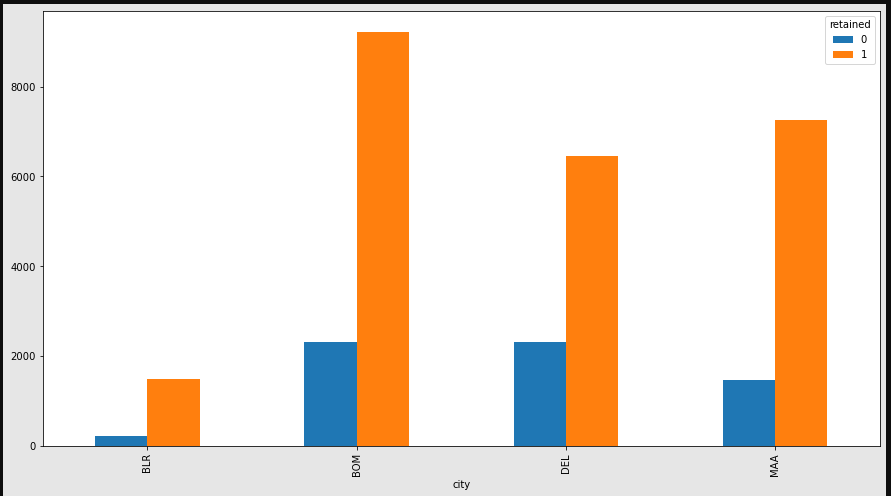
We have plotted **a bar plot on favday** to understand the trends of delivery and orders across the different days of the week.



### Inferences:

* The number of orders delivered are highest on Monday and Tuesday since the number of orders placed are maximum on Saturday and Sunday.

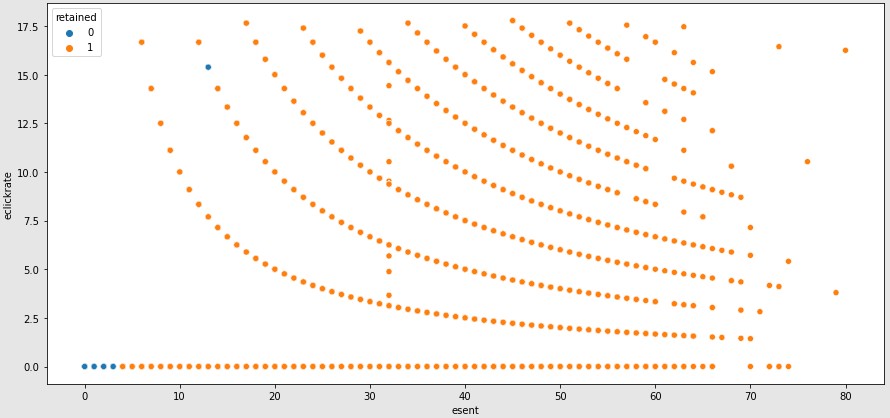
We have plotted a **bar plot on orders across different cities** to find the retention rate in different cities

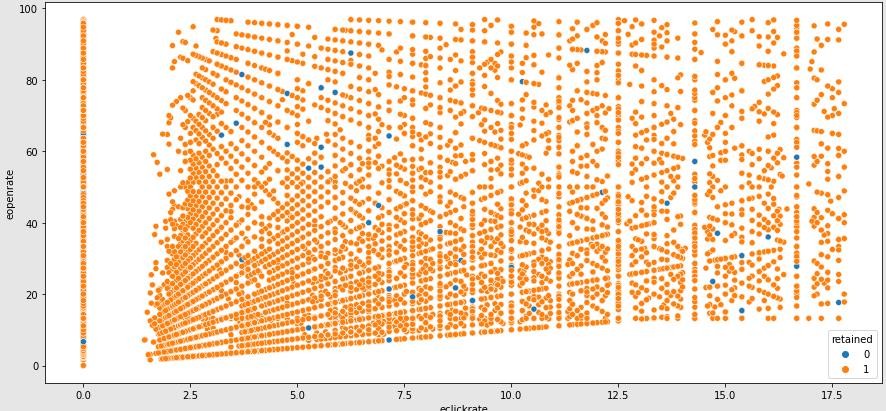


**Inferences -**

* The retention rate is maximum in Bombay because of high population and rainy climatic conditions.
* Bangalore has the least retention rate because people in Bangalore prefer coffee over tea.

We have plotted **a scatter plot** for the fields **eclickrate** and **esent** to understand the retention trend.

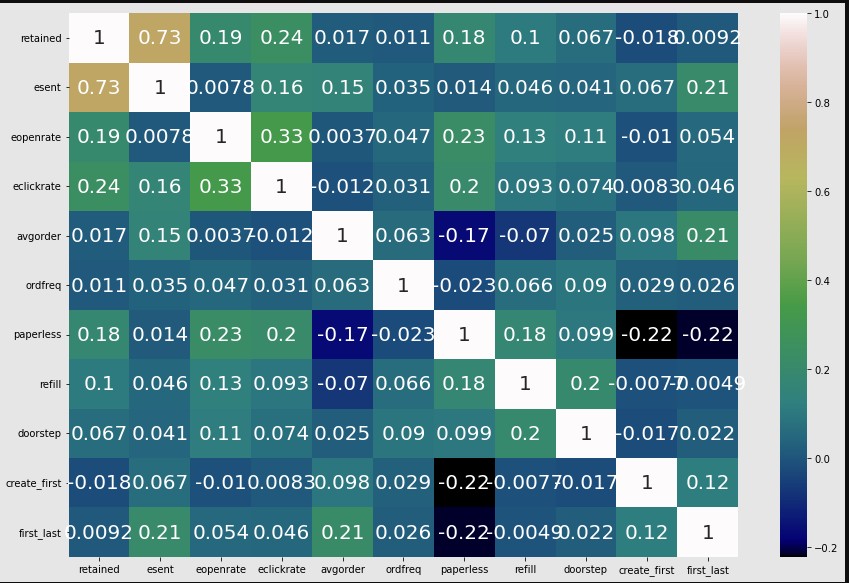




### Inferences -

* As esent increases customers are likely to be retained.

We plotted a heatmap to check the correlation



### Inferences -

* Positive correlation for esent and retained
* Positive correlation(moderate) for esent and eopenrate
* Paperless and Avg order are negatively correlated

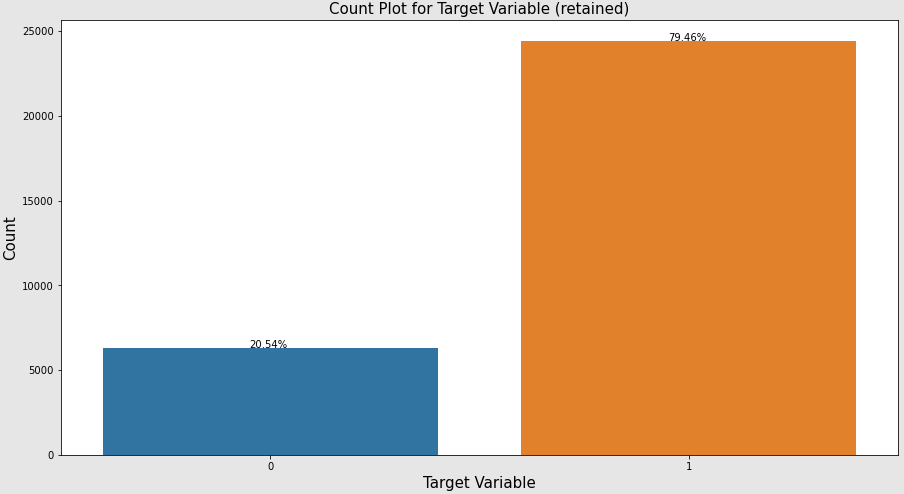
We have used transformation techniques like boxcox and we have scaled the data using standard scaler method.

#### Transformation techniques -

We have observed that few of our columns such as eopen rate,eclickrate ,avgorder and order frequency are highly right skewed hence we are using boxcox to normalize the data.

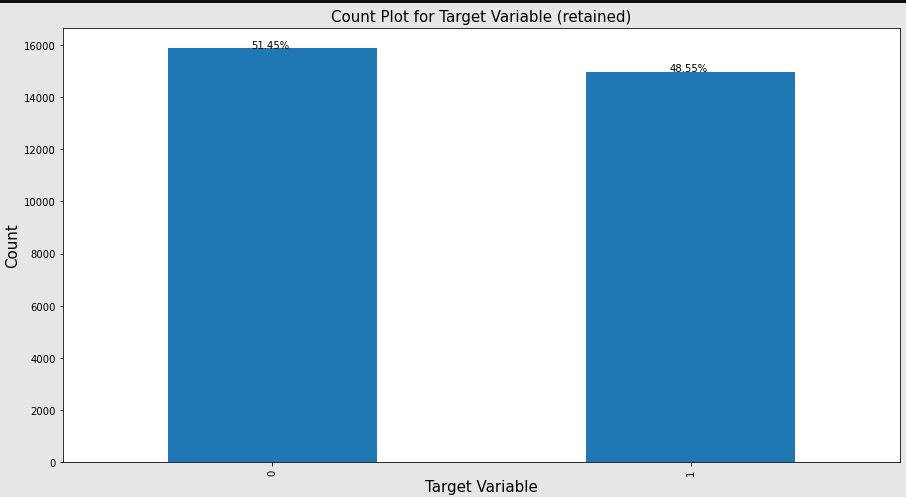
We have used the standard scalar to scale the data. Scaling makes it easy for the model to learn and understand the problem.

**Class imbalance and treatment:**



We can infer that the data is imbalanced.

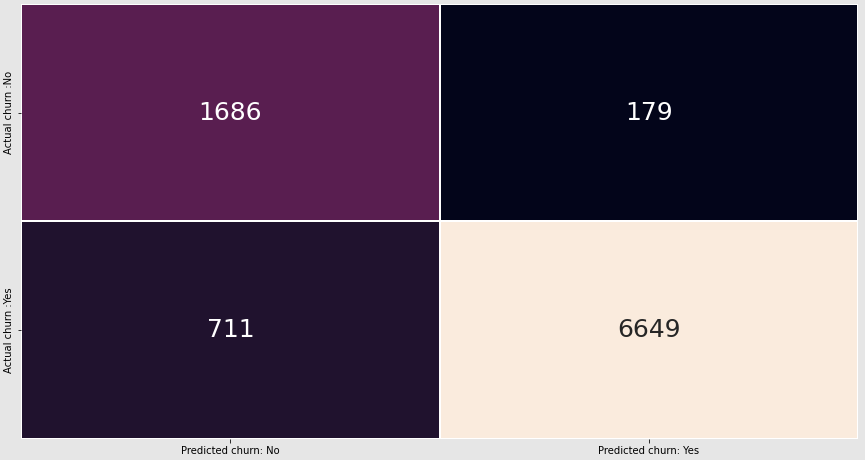
After Treating the imbalanced data, we can now observe that the data is balanced.



# Base Model

We have used logistic regression and we have fitted out training data. We have predicted the ypred and we have used list comprehension by assuming that threshold value as (>0.5 - Retained & <0.5 - Not Retained).

We have plotted a confusion matrix to check whether our model has predicted the churn rate.



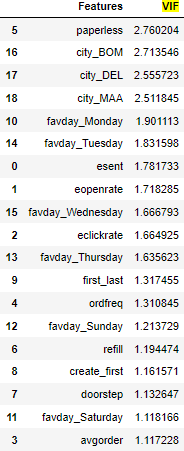
By using the classification report we can find the churn precision, churn recall and Churn f1 score.

To find the accuracy score and Roc\_Auc\_score we are using a function from sklearn metrics.



# VIF (Variance Inflation Factor)

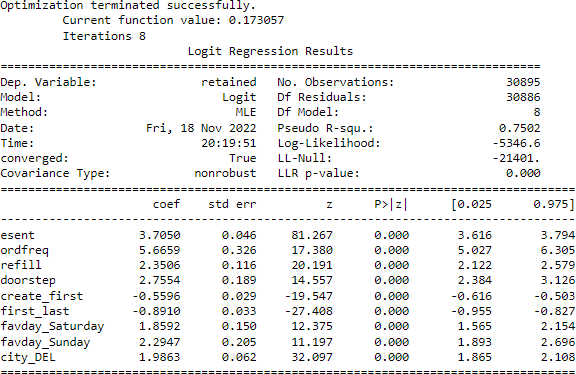
To check Multi collinearity



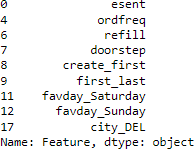
#### As the value of every feature is less than the 5 we can say that the presence of multicollinearity is not present.

**Feature Selection**

RFE is popular because it is easy to configure and use. It is effective in selecting those features in a training dataset that are most relevant in predicting the target variable.

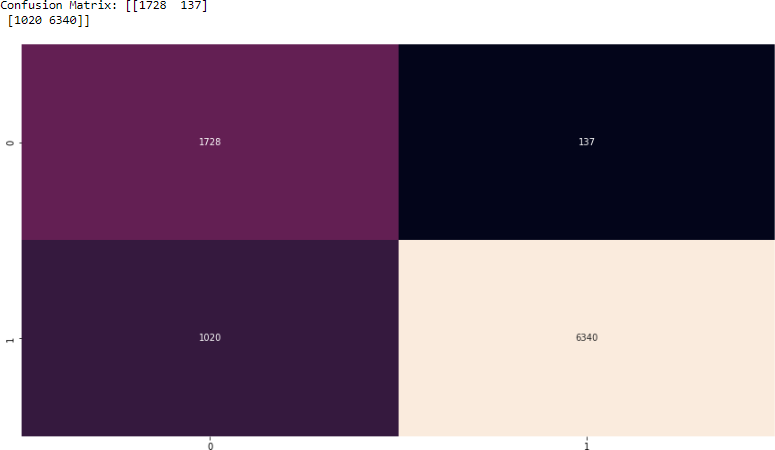


Using the Feature Selection RFE Model the features which are affecting the target variable are the following



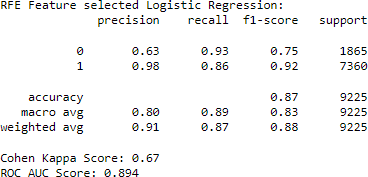
**Confusion Matrix for RFE model:**

We have plotted a confusion matrix to check whether our model has predicted the churn rate.

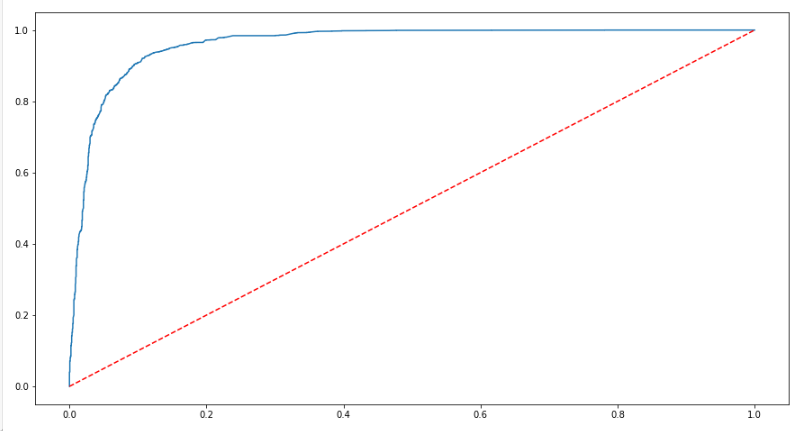


By using the classification report we can find the churn precision, churn recall and Churn f1 score.

To find the accuracy score and Roc\_Auc\_score we are using a function from sklearn metrics.



ROC AUC Curve for RFE model:



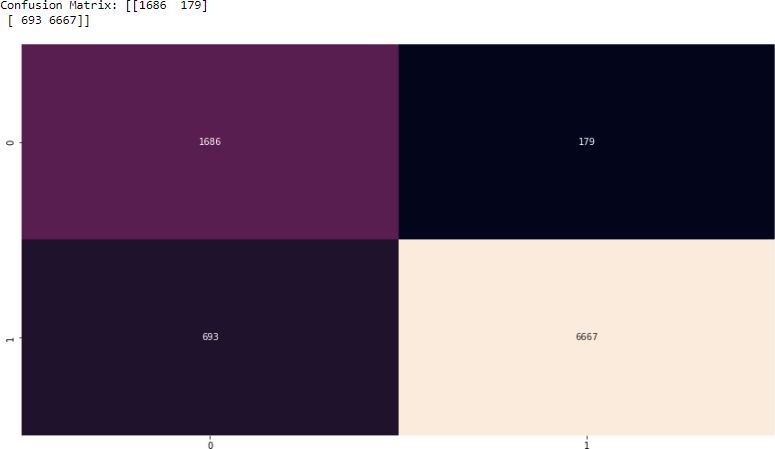
# Hyper parameter tuning:

**Decision Tree**

In Decision Tree we are using hyper parameter tuning using GridSearchCV & we are fitting the training data to get the best parameters.

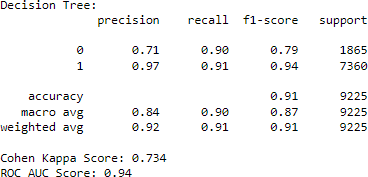
We have tuned the model using those parameters and fitted the model.Post that we have predicted the ypred.

We have plotted a confusion matrix to check whether our model has predicted the churn rate.

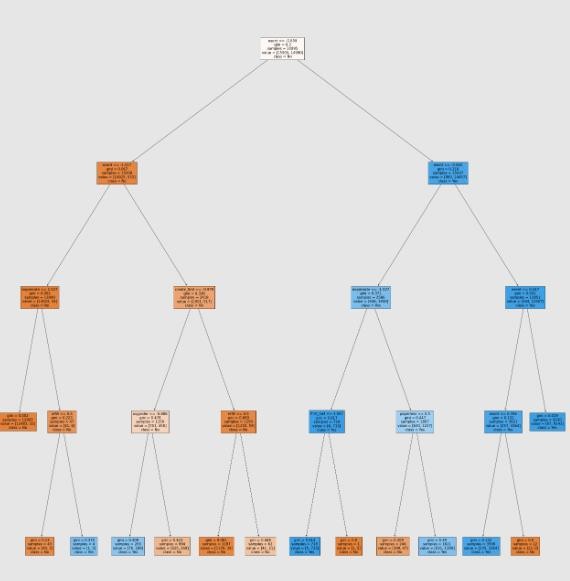


By using the classification report we can find the churn precision, churn recall and Churn f1 score.

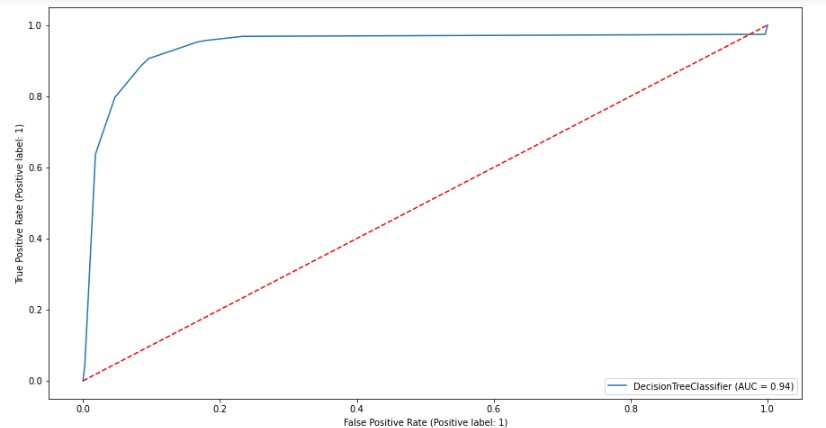
To find the accuracy score and Roc\_Auc\_score we are using a function from sklearn metrics.



We have plotted the decision tree



ROC AUC Curve for decision tree.



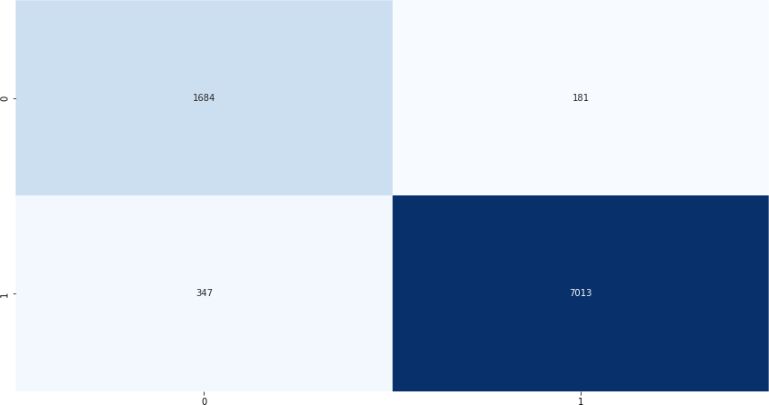
ROC\_AUC score for decision tree is 0.905.

# Random Forest

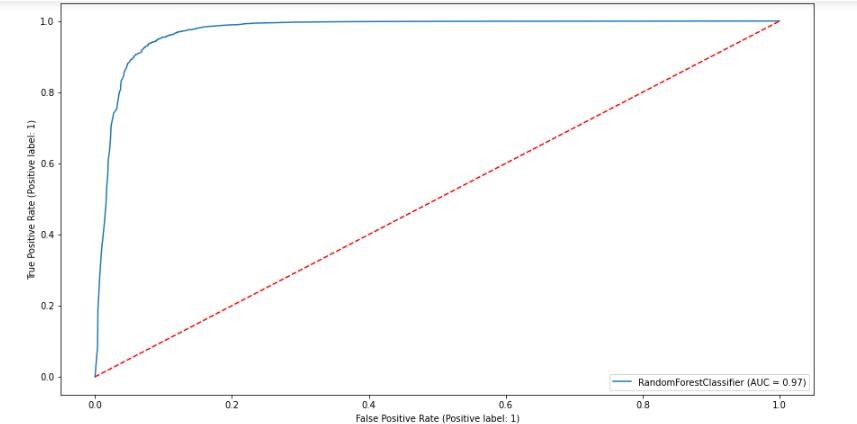
In Random Forest we are using hyper parameter tuning using GridSearchCV & we are fitting the training data to get the best parameters.

We have tuned the model using those parameters and fitted the model.Post that we have predicted the ypred.

We have plotted a confusion matrix to check whether our model has predicted the churn rate.



ROC AUC Curve for Random Forest ROC\_AUC score for Random Forest is 0.928



# Boosting Techniques:

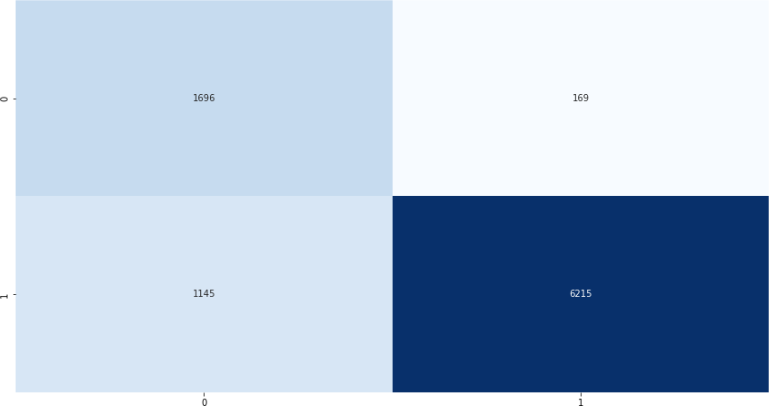
We have used two boosting methods. 1.Adaboost

1. XGboost

# Adaboost

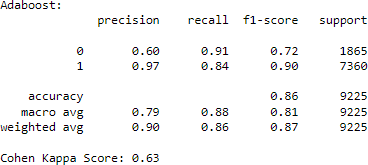
In Adaboost we are using base estimator as decision tree classifier and fitting the X train, Y train and finding the ypred.

We have plotted a confusion matrix to check whether our model has predicted the churn rate.

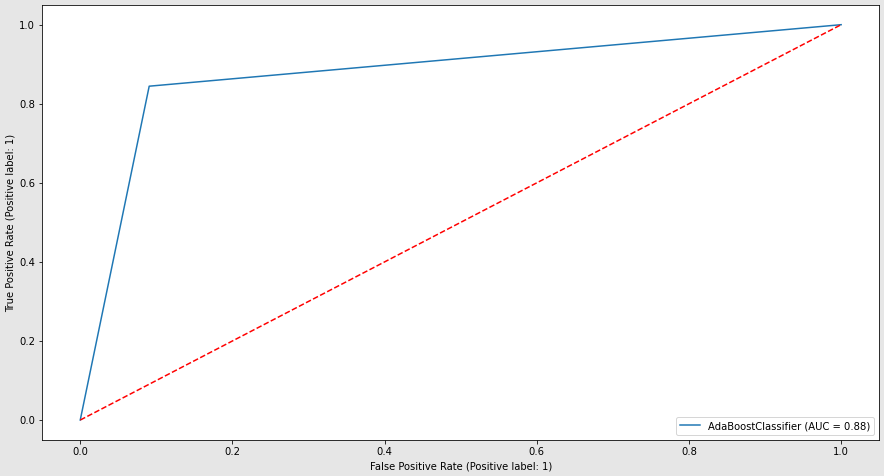


By using the classification report we can find the churn precision, churn recall and Churn f1 score.

To find the accuracy score and Roc\_Auc\_score we are using a function from sklearn metrics.



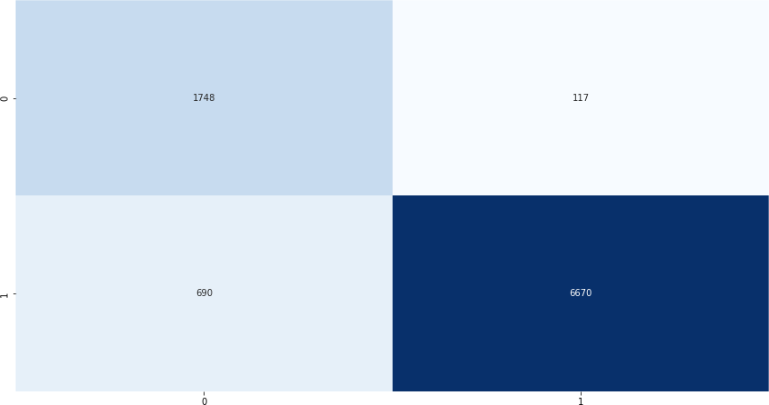
ROC AUC Curve for Adaboost ROC\_AUC score for Adaboost is 0.877



# XGboost

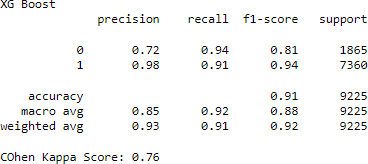
In XGboost we are initiating the XGF classifier AND fitting the X train,Y train.Post that we find the ypred.

We have plotted a confusion matrix to check whether our model has predicted the churn rate.



By using the classification report we can find the churn precision, churn recall and Churn f1 score.

To find the accuracy score and Roc\_Auc\_score we are using a function from sklearn metrics.

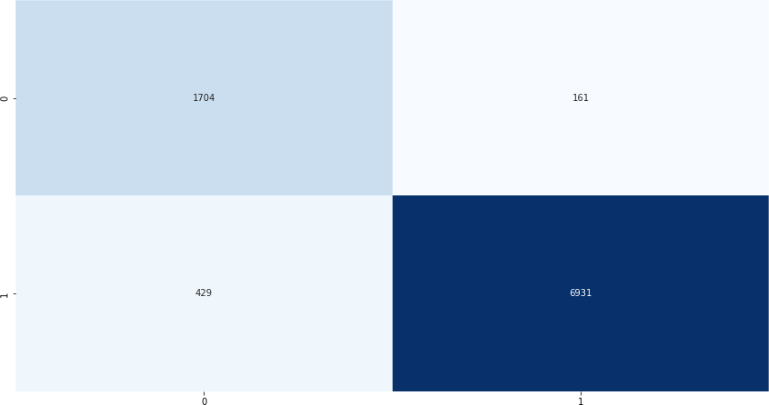


ROC\_AUC score for XGboost is 0.92

**Stacking**

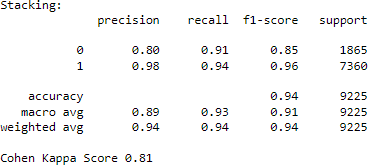
In stacking we have used the base as decision tree classifier and random forest classified and Ada boost classifier as the final estimator. Post that we find the ypred.

We have plotted a confusion matrix to check whether our model has predicted the churn rate.



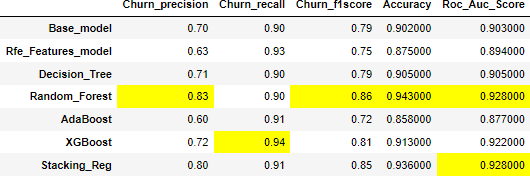
By using the classification report we can find the churn precision, churn recall and Churn f1 score.

To find the accuracy score and Roc\_Auc\_score we are using a function from sklearn metrics.



ROC AUC Curve for Stacking ROC\_AUC score for Stacking is 0.928

# Metric Table



# Comparison & Selection of Model:

The accuracy of the Random\_Forest model is comparatively better than other models. Accuracy of Random Forest Model - 94.3% The Stacking model gave a similar accuracy of 93.6% but since the random forest had a better precision rate and f1 score, we have selected the Random\_Forest model.

# Results and Discussion:

1. Based on the results it is evident that our dataset performs better with the Random\_Forest Model.
2. The accuracy of the Random\_Forest model is comparatively better than other models. Accuracy of Random Forest Model - 94.3%
3. Features affecting our target variable - esent, favday\_Sunday, favday\_Saturday, city\_DEL, create\_first, doorstep, first\_last, ordfreq, refill.

# Limitations:

1. Large investment in terms of price and time: Customer retention can prove expensive for business in the way that it involves large investment both in terms of price and time.
2. It requires huge cost for running loyalty programs in order to retain customers for a longer period. Business needs to sacrifice their profit by offering several discount and cashback offers to its audience.
3. Require concerted commitment and Business Culture: Every business organization for attaining better customer retention rate should ensure a concerted commitment and proper culture. Every member working at distinct hierarchy of organization should focuses on providing best services to their customers.
4. New customers may be overlooked: Organization making efforts for attaining efficient customer retention rate may not focus on needs of new customers. There may be chances of new clients being overlooked by brand in a hoard to satisfy its existing customers. Unsatisfied customers may spread negative piece of information about the brand in society.
5. The dataset includes data from the year 2008-2018.The First five years 2008- 2013 the customer behavior towards online shopping would have been different as ecommerce had not boomed by then.
6. The Customer behavior can be unpredictable.

# Conclusion:

1. Targeted Email campaign for the existing customers will increase customer retention rate.
2. Faster 2-day delivery is an effective way to increase customer retention rate.
3. By using the Random Forest model we can predict the churn rate with an accuracy of 94.3%.

# Recommendations:

1. Strong Customer service strategy to create a memorable experience to ensure repeat purchase.
2. Be a part of the social conversation.
3. Give gifts, rewards and prizes that are personalized and relevant to their interests
4. Create exclusive offers, hold special events, grant them early access to new offerings and VIP services.
5. Involve them in your new product planning processes.
6. Feature them in case studies and testimonials and podcasts.
7. Create a loyalty program.

# Assumptions to be satisfied:

* The logistic regression assumes that there is minimal or no multicollinearity among the independent variables.
* The Logistic regression assumes that the independent variables are linearly related to the log of odds.
* Logistic regression usually requires a large sample size to predict properly.
* Logistic regression assumes that the dependent variable is binary
* The Logistic regression assumes the observations to be independent of each other.
* We have assumed that the threshold value to be 0.5 in the base model.

# References:

1. Ascarza, E. (2018). Retention futility: Targeting high-risk customers might be ineffective. Journal of Marketing Research, 55(1), 80–98[. https://doi.org/10.1509/jmr.16.0163](https://doi.org/10.1509/jmr.16.0163)
2. Abu Zayyad, H.M., Obeidat, Z.M., Alshurideh, M.T., Abuhashesh, M., Maqableh, M., Masa’deh, R.E.: Corporate social responsibility and patronage intentions: the mediating effect of brand credibility. J. Mark. Commun. 1–24 (2020)
3. Alshurideh, D.M.: Do electronic loyalty programs still drive customer choice and repeat purchase behaviour? Int. J. Electron. Cust. Relatsh. Manag. 12(1), 40–57 (2019)
4. Zeithaml, V.A., Berry, L.L., Parasuraman, A.: The behavioral consequences of service quality. J. Mark. 60(2), 31–46 (1996)
5. Danesh, S.N., Nasab, S.A., Ling, K.C.: The study of customer satisfaction, customer trust and switching barriers on customer retention in Malaysia hypermarkets. Int. J. Bus. Manag. 7 (7), 141–150 (2012)
6. Al-Dmour, H., Alshuraideh, M., Salehih, S.: A study of Jordanians’ television viewers habits. Life Sci. J. 11(6), 161–171 (2014)
7. Alshurideh, M.: The factors predicting students’ satisfaction with universities’ healthcare clinics’ services: a case-study from the Jordanian higher education sector. Dirasat. Adm. Sci. 161(1524), 1–36 (2014)
8. Preikschas, M.W., Cabanelas, P., Rüdiger, K., Lampón, J.F.: Value co-creation, dynamic capabilities and customer retention in industrial markets. J. Bus. Ind. Mark. 32(3), 409–420 (2017)
9. Al Dmour, H., Alshurideh, M., Shishan, F.: The influence of mobile application quality and attributes on the continuance intention of mobile shopping. Life Sci. J. 11(10), 172–181 (2014)
10. Obeidat, B., Sweis, R., Zyod, D., Alshurideh, M.: The effect of perceived service quality on customer loyalty in internet service providers in Jordan. J. Manag. Res. 4(4), 224–242 (2012)
11. Al-dweeri, R., Obeidat, Z., Al-dwiry, M., Alshurideh, M., Alhorani, A.: The impact of eservice quality and e-loyalty on online shopping: moderating effect of e-satisfaction and etrust. Int. J. Mark. Stud. 9(2), 92–103 (2017).
12. Levesque, T., McDougall, G.H.G.: Determinants of customer satisfaction in retail banking. Int. J. bank Mark. 14(7), 12–20 (1996)
13. Hansemark, O.C., Albinsson, M.: Customer satisfaction and retention: the experiences of individual employees. Manag. Serv. Qual. Int. J. 14(1), 40–57 (2004)
14. Morgan, R.M., Hunt, S.D.: The commitment-trust theory of relationship marketing. J. Mark. 58(3), 20–38 (1994).

s