In [168]:
```
import warnings
warnings.filterwarnings('ignore')

import pandas as pd
import numpy as np
from plotnine import *
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.cluster import AgglomerativeClustering
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeClassifier

from sklearn.cluster import KMeans
from sklearn.mixture import GaussianMixture
from sklearn.model_selection import LeaveOneOut
from sklearn.naive_bayes import GaussianNB, BernoulliNB, MultinomialNB,
CategoricalNB
from sklearn.metrics import silhouette_score
from sklearn.preprocessing import LabelBinarizer
import scipy.cluster.hierarchy as sch
from matplotlib import pyplot as plt
from sklearn.metrics import plot_confusion_matrix
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn.metrics import silhouette_score
import pandas as pd
import numpy as np
from plotnine import *
import statsmodels.api as sm
import statsmodels.formula.api as smf

%matplotlib inline
```

Loading the Data

In [184]:
```
champion = "/Users/ishansupanekar/Documents/MGSC410/MGSC410/MasterJshan_v2.csv"
championdf = pd.read_csv(champion)
championdf.head()#read the csv file (put 'r' before the path string to address any special characters in the path, such as '\'). Don't forget to put the file name at the end of the path + ".csv"
championdf.shape
championdf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 38609 entries, 0 to 38608
Data columns (total 36 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   ID                              38609 non-null  int64  
 1   Puchase_Amounts                 25430 non-null  float64
 2   Duration                        38609 non-null  int64  
 3   Send_Count                      38609 non-null  int64  
 4   Open_Count                      38609 non-null  int64  
 5   Open_Rate                       38609 non-null  float64
 6   Unique_Click_Count              38609 non-null  int64  
 7   start_count                     38609 non-null  int64  
 8   other_count                     38609 non-null  int64  
 9   completed_count                 38609 non-null  int64  
 10  onboarding_count                38609 non-null  int64  
 11  app_launch_count                38609 non-null  int64  
 12  champion_score                  38609 non-null  float64
 13  champion_binary                 38609 non-null  int64  
 14  Language_ALL                    38609 non-null  int64  
 15  Language_ENG                    38609 non-null  int64  
 16  Language_ESP                    38609 non-null  int64  
 17  Language_FRA                    38609 non-null  int64  
 18  Language_ITA                    38609 non-null  int64  
 19  Language_Other                  38609 non-null  int64  
 20  Subscription_Type_Limited_Yes   38609 non-null  int64  
 21  Subscription_Type_Lifetime_Yes  38609 non-null  int64  
 22  Subscription_Event_Type_RENEWAL 38609 non-null  int64  
 23  Purchase_Store_Web_Yes          38609 non-null  int64  
 24  Demo_User_Yes                   38609 non-null  int64  
 25  Free_Trial_User_Yes             38609 non-null  int64  
 26  Auto_Renew_On                   38609 non-null  int64  
 27  Country_Europe                  38609 non-null  int64  
 28  Country_Other                   38609 non-null  int64  
 29  Country_US/Canada               38609 non-null  int64  
 30  User_Type_Consumer              38609 non-null  int64  
 31  Lead_Platform_App               38609 non-null  int64  
 32  Lead_Platform_Unknown           38609 non-null  int64  
 33  Lead_Platform_Web               38609 non-null  int64  
 34  Email_Subscriber_Yes            38609 non-null  int64  
 35  Push_Notifications_Yes          38609 non-null  int64  
dtypes: float64(3), int64(33)
memory usage: 10.6 MB
```

Testing and Training on predictors

In [274]:
```
predictors = ["start_count","Purchase_Store_Web_Yes","Demo_User_Yes","Free_Trial_User_Yes","Country_US/Canada","Lead_Platform_App","Auto_Renew_On","Subscription_Type_Limited_Yes","User_Type_Consumer","Email_Subscriber_Yes"]
X_train, X_test, y_train, y_test = train_test_split(championdf[predictors], championdf["champion_binary"], test_size=0.2)
```

In [275]:
```
X_train.shape
```

Out[275]:
```
(30887, 10)
```

In [276]:
```
X_test.shape
```

Out[276]:
```
(7722, 10)
```

In [277]:
```
X_train.head()
```

Out[277]:

	start_count	Purchase_Store_Web_Yes	Demo_User_Yes	Free_Trial_User_Yes	Country_US/Ca
17515	0	0	0	1	
13426	2	0	1	0	
27575	0	1	0	0	
2819	3	1	1	0	
24943	0	0	1	0	

Logit Model with accuracy scores

In [278]:
```
myLogit = LogisticRegression()
```

In [279]:
```
logit = myLogit.fit(X_train,y_train)
```

In [280]:
```
predictedVals = myLogit.predict(X_test)
```

In [281]:
```
accuracy_score(y_test,predictedVals)
```

Out[281]:
```
0.941983941983942
```

In [282]:
```
confusion_matrix(y_test,predictedVals)
```

Out[282]:
```
array([[7248,   41],
       [ 407,   26]])
```

In [283]:
```
print("Training set score: {:.3f}".format(logit.score(X_train,y_train)))
print("Test set score: {:.3f}".format(logit.score(X_test,y_test)))
```

```
Training set score: 0.946
Test set score: 0.942
```

In [284]:
```
logit_model=sm.Logit(y_train,X_train)
result=logit_model.fit()
print(result.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.193434
         Iterations 8
                           Logit Regression Results
==============================================================================
=======
Dep. Variable:        champion_binary   No. Observations:
30887
Model:                          Logit   Df Residuals:
30877
Method:                           MLE   Df Model:
9
Date:                Wed, 09 Dec 2020   Pseudo R-squ.:
0.07949
Time:                        18:13:03   Log-Likelihood:
-5974.6
converged:                       True   LL-Null:
-6490.5
Covariance Type:            nonrobust   LLR p-value:                2.2
76e-216
==============================================================================
============================
                                   coef    std err          z      P>|
z|      [0.025      0.975]
------------------------------------------------------------------------------
--------------------------
start_count                      0.0629      0.002     31.876      0.0
00       0.059       0.067
Purchase_Store_Web_Yes          -1.8780      0.045    -41.481      0.0
00      -1.967      -1.789
Demo_User_Yes                    0.1777      0.095      1.869      0.0
62      -0.009       0.364
Free_Trial_User_Yes              0.4759      0.082      5.829      0.0
00       0.316       0.636
Country_US/Canada               -0.4155      0.069     -5.981      0.0
00      -0.552      -0.279
Lead_Platform_App               -1.5702      0.089    -17.660      0.0
00      -1.745      -1.396
Auto_Renew_On                    0.6878      0.067     10.240      0.0
00       0.556       0.819
Subscription_Type_Limited_Yes   -2.6204      0.060    -43.734      0.0
00      -2.738      -2.503
User_Type_Consumer               0.5085      0.087      5.872      0.0
00       0.339       0.678
Email_Subscriber_Yes            -0.0179      0.072     -0.250      0.8
03      -0.158       0.122
==============================================================================
============================
```

Output coefficients in Odds Ratios

In [285]:
```
np.exp(result.params)
```

Out[285]:
```
start_count                      1.064903
Purchase_Store_Web_Yes           0.152896
Demo_User_Yes                    1.194506
Free_Trial_User_Yes              1.609534
Country_US/Canada                0.660036
Lead_Platform_App                0.207993
Auto_Renew_On                    1.989415
Subscription_Type_Limited_Yes    0.072775
User_Type_Consumer               1.662821
Email_Subscriber_Yes             0.982294
dtype: float64
```

In [286]:
```
pd.set_option('display.max_rows', None)
```

In [ ]:

In [ ]:

In [ ]: