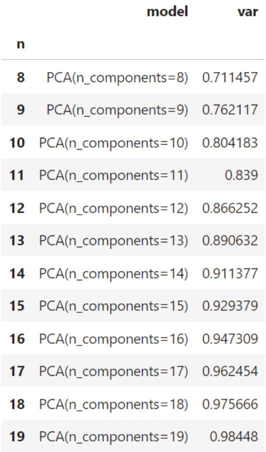
This report will be analyzing preprocessed customer churn data to identify patterns behind churned customers by focusing on clustering of the churned customer dataset. The result of this analysis should provide more insight into reasons why certain customers would churn and help improve the sales model.

Preprocessing

The customer churn data consist of total 23 features including preprocessed categorical features via one hot encoding as integers and true/false churned value. For the purpose of this analysis, only churned customers were selected as the sample dataset (1869 / 7043). Non-categorical features were then analyzed for potential skew. Only gb_mon feature was identified with a mild skew and was normalized using log transformation followed by min-max scaling.

PCA was then performed on the churned dataset to reduce the dimensionality while retaining ~90% of the variance at 14 components.

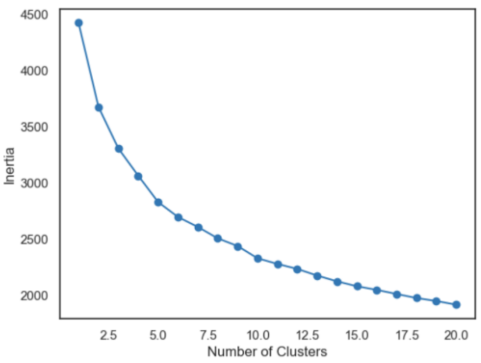


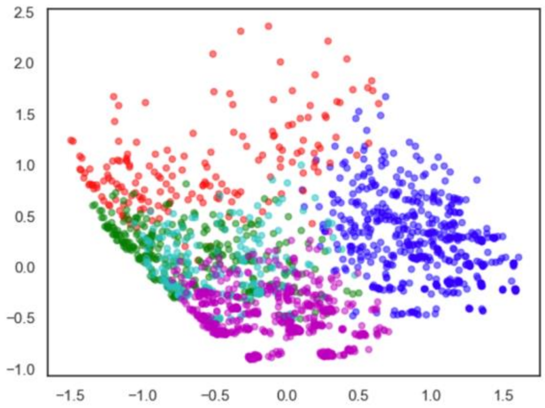
n	model	var
8	PCA(n_components=8)	0.711457
9	PCA(n_components=9)	0.762117
10	PCA(n_components=10)	0.804183
11	PCA(n_components=11)	0.839
12	PCA(n_components=12)	0.866252
13	PCA(n_components=13)	0.890632
14	PCA(n_components=14)	0.911377
15	PCA(n_components=15)	0.929379
16	PCA(n_components=16)	0.947309
17	PCA(n_components=17)	0.962454
18	PCA(n_components=18)	0.975666
19	PCA(n_components=19)	0.98448

Clustering Model

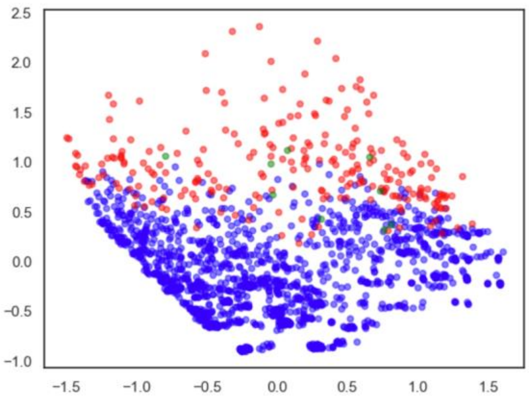
The dataset was then subjected to 3 clustering methods.

K-means clustering was performed as base using n = 5 selected empirically based on the elbow method ranging from n = 1-20 on inertia resulting in 5 clusters shown in the graph.

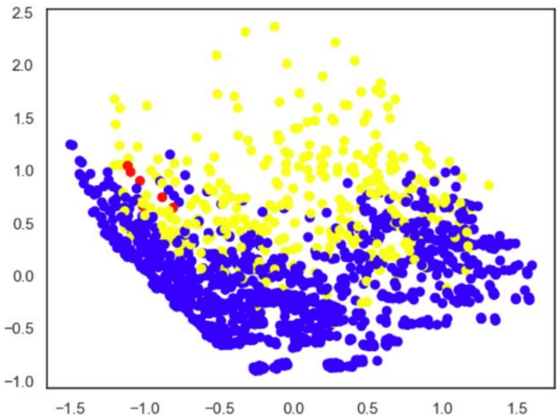




Mean shift was performed on the same dataset where bandwidth was chosen from the estimate method using 10% quantile resulting in 3 clusters



Finally, DBSCAN was performed on the same dataset using 0.5% of the sample size (n=9) as the minimum size of the core and eps = 1 as the maximum distance resulting in 2 clusters and 320/1869 outliers.



Insights

Overall, both DBSCAN and mean shift method seem to show similar number and distribution of clusters while k-mean had more clusters identified. Perhaps the size of clusters are uneven and k-means was unable to group certain datapoints together. Since the details are unknown at this point, both mean shift and DBSCAN seem to be a good starting point for the further analysis at identifying number of clusters and potential outliers in the dataset.

For further analysis, subjecting the dataset to ward method may yield more information on the cluster size and hierarchy, which could explain the performance of k-means. The models could be more fine tuned with different set of hyperparameters as well for further insights. Once certain level of confidence is achieved, comparing the model with non-churned customer data to measure significance of the findings would help determining the validity of the cluster models as well.