Toronto Metropolitan University — Data Science & Analytics Graduate Studies

2022-2023 M.Sc. in Data Science and Analytics

Improving Customer Churn Prediction Through Segmentation

Student Name: Smit Patel **Supervisor: Dr. Ozgur Turetken**

Objective

The objective of this research is to determine if the accuracy of the existing churn prediction models can be improved through the use of clustering customer data using segmentation techniques as opposed to feeding an entire dataset into a churn prediction model. Using a combinations of 3 different churn prediction models (with one involving neural networks) along with 3 different customer segmentation models will help identify which combination yields the highest test accuracy prediction.

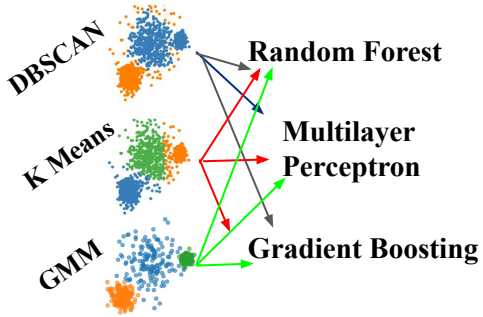
Background

A large issue companies face is seeing a ROI on initial high capital marketing investments. A study found that acquiring a new customer is anywhere from 5-25% times more expensive than retaining an existing one and another found that increasing customer retention rates by just 5% increases profits from anywhere between 25-95%

Literature review for both areas found that the top clustering methods for segmentation were KMeans, DBSCAN and GMM while the top existing churn prediction models which yielded highest test accuracy were Random Forest and Gradient Boosting.

Methodology

1. Preprocess data by removing outliers, balancing target variable using SMOTE, identify important features using Random Forest and use PCA to create subset dataset
2. Segment data using each clustering method
3. Fit training data from each individual cluster group to each churn prediction model and take weighted average between all clusters
4. Compare testing accuracy between each combination of segmentation + prediction method



Results

Churn Prediction - No Segmentation

	Model	Train Accuracy	Test Accuracy	Test f1-Score
0	Random Forest	1.000000	0.822077	0.815382
1	MLP	0.987282	0.809350	0.808776
2	Gradient Boosting	1.000000	0.835252	0.833922

Churn Prediction - K Means Segmentation

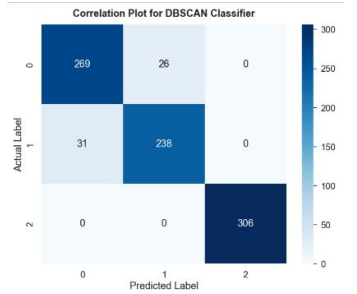
	Prediction Model	Training Accuracy (%)	Testing Accuracy (%)	Testing f1-score (%)
0	Random Forest	1.000000	0.870050	0.860888
1	MLP	1.000000	0.804455	0.804899
2	Gradient Boosting	1.000000	0.847772	0.844289

Churn Prediction - DBSCAN Segmentation

	Prediction Model	Training Accuracy (%)	Testing Accuracy (%)	Testing f1-Score (%)
0	Random Forest	1.000000	0.925287	0.922891
1	MLP	1.000000	0.920690	0.918279
2	Gradient Boosting	1.000000	0.934483	0.932420

Churn Prediction - GMM Segmentation

	Prediction Model	Training Accuracy (%)	Testing Accuracy (%)	Testing f1-score (%)
0	Random Forest	1.000000	0.851613	0.841816
1	MLP	0.239748	0.266276	0.153665
2	Gradient Boosting	1.000000	0.845161	0.843894



Conclusions

The combination of DBSCAN clustering along with Gradient Boosting yields the highest churn prediction accuracy of 93.45%, which is higher than the churn prediction when no segmentation was conducted. Thus, segmenting data improves churn prediction accuracy!