**Customer Churn Analysis**

Weekly Report

Ahmedabad University

4rier Series

Prof. Mehul Raval

CSE523 - Machine Learning

**Balancing unbalanced data using sampling**

Original Data:

Counter({0: 5163, 1: 1869})

1. Random Oversampling

ros = RandomOverSampler()

# resampling X, y

X\_ros, y\_ros = ros.fit\_resample(X, y)

# new class distribution

counter = Counter(y\_ros)

print(counter)

Counter({0: 5163, 1: 5163})

1. Random Undersampling

rus = RandomUnderSampler()

# resampling X, y

X\_rus, y\_rus = rus.fit\_resample(X, y)

# new class distribution

print(Counter(y\_rus))

Counter({0: 1869, 1: 1869})

1. Combining Undersampling and Oversampling

over = RandomOverSampler(sampling\_strategy=0.5)

under = RandomUnderSampler(sampling\_strategy=0.8)

# first performing oversampling to minority class

X\_over, y\_over = over.fit\_resample(X, y)

print(f"Oversampled: {Counter(y\_over)}")

# now to combine under sampling

X\_combined\_sampling, y\_combined\_sampling = under.fit\_resample(X\_over, y\_over)

print(f"Combined Random Sampling: {Counter(y\_combined\_sampling)}")

Oversampled: Counter({0: 5163, 1: 2581})

Combined Random Sampling: Counter({0: 3226, 1: 2581})

1. SMOTE Sampling

from imblearn.over\_sampling import SMOTE

smoteSampler = SMOTE()

X\_smote, y\_smote = smoteSampler.fit\_resample(X, y)

# summarize the new class distribution

counter = Counter(y\_smote)

print(counter)

Counter({0: 5163, 1: 5163})

On the part of analysis taken before and after, we found some surprising and unusual results.

Statistics when all 31 features, when taken into consideration:

| Factors | Logistic Regression | SVM | Naive Bayes |
| --- | --- | --- | --- |
| Accuracy | 81.27962085308057 % | 80.52132701421802 % | 65.260663507109 % |
| Precision | 66.04166666666667 % | 66.42857142857143 % | 41.91564147627416 % |
| Recall | 57.74134790528234 % | 50.81967213114754 % | 86.88524590163934 % |
| F1 score | 80.85436799018186 % | 79.61121682882954 % | 67.2861386096018 % |

Statistics when only 12 features are taken into consideration using pearson correlation by setting threshold limit:

| factors | Logistic Regression | SVM | Naive Bayes |
| --- | --- | --- | --- |
| Accuracy | 80.4739336492891 % | 79.81042654028437 % | 76.49289099526067 % |
| Precision | 65.46275395033861 % | 66.2269129287599 % | 54.04580152671755 % |
| Recall | 52.82331511839708 % | 45.71948998178507 % | 64.48087431693989 % |
| F1 score | 79.75130157938099 % | 78.48246426136785 % | 77.11458325458942 % |

Statistics when SMOTE analysis are taken into consideration:

| factors | Logistic Regression | SVM | Naive Bayes |
| --- | --- | --- | --- |
| Accuracy | 81.56398104265404 % | 73.98104265402844 % | 65.260663507109 % |
| Precision | 80.86344832844314 % | 54.73194672177175 % | 79.40697829739885 % |
| Recall | 81.56398104265404 % | 73.98104265402844 % | 65.260663507109 % |
| F1 score | 81.07683505511926 % | 62.91713842709802 % | 67.2861386096018 % |

Though accuracy. Precision, f1 score and recall have increased in the case of logistic regression compared to when features were selected on the basis of threshold using our function. But there in some unusualities along with it. The confusion matrix for SVM shows false and true positives as 0 which means there are chances of data imbalance. And we are figuring out the problem to fix it.