**Assignment**

**Wali Ullah (09745)**

**Stroke Data**

The stroke dataset has 5110 rows and 12 columns. The cleaning function in template is called to change columns to numeric. The bmi column has missing values which are filled by linear interpolation and un-necessary columns are deleted.

* The heatmap of correlation shows that the correlation among features is not so high.
* t-stats for equality of mean of all numeric columns suggest that mean of numeric columns are not equal
* the Chi-square statistics suggest that numeric columns are not linearly independent
* The data of numeric columns is not normal.

The one hot encoding is applied to string columns to create dummies, which results in

Rows=5110, and columns= 22)

Machine Learning algorithms are applied in two stages.

1. In the first sage the issue of class imbalance is not addressed. Eleven different machine learning classification algorithms are applied and four different scenarios are considered
   1. without focusing the cross validation (CV) and features selection.
   2. Only cross validation (CV: Stratified K-Fold) is considered
   3. Only feature selection (random forest based algorithm is used) criterion is considered
   4. Both cross validation (CV: Stratified K-Fold) and features selection (random forest based algorithm is used) are considered.

The results of this stage indicate that

* With the application of CV the precision and accuracy increase in comparison to the bench mark category which is (without CV and RFFS)
* The performance in terms all four criterions (precision, recall, AUC and accuracy) worsens with the application of features selection criterion in comparison to the benchmark category
* The final Selected Features are ['avg\_glucose\_level', 'age', 'bmi', 'hypertension']

Indicating that categorical variables do not play any role.

* With the application of both CV and RFFS criterions, almost all 11 algorithms outperform the benchmark as well as the other two categories.
* The best performing algorithm seems to be the Random Forest. The performance of all remaining algorithms is considerably worse.

1. In the second stage the class imbalance issue is resolved through SMOT oversampling (to increase the under sampled category which is stroke=1). The results show that

Original data shape: (5110, 22)

Resampled data shape: (9722, 22)

The above four different scenarios are considered

1. without focusing the cross validation (CV) and features selection.
2. Only cross validation (CV: Stratified K-Fold) is considered
3. Only feature selection (random forest based algorithm is used) criterion is considered
4. Both cross validation (CV: Stratified K-Fold) and features selection (random forest based algorithm is used) are considered.

The results of this stage indicate that

* With the application of CV the precision and accuracy increase in comparison to the bench mark category which is (without CV and RFFS)
* The performance in terms all four criterions (precision, recall, AUC and accuracy) worsens a little bit with the application of features selection criterion in comparison to the benchmark category
* Selected Features are =['age', 'avg\_glucose\_level', 'smoking\_status\_Unknown', 'smoking\_status\_never smoked', 'bmi', 'smoking\_status\_formerly smoked', 'Residence\_type\_Urban', 'ever\_married\_No', 'Residence\_type\_Rural', 'smoking\_status\_smokes', 'work\_type\_Private', 'work\_type\_Govt\_job', 'work\_type\_Self-employed']
* With the application of both CV and RFFS criterions, almost all 11 algorithms underperform than the CV only based scenario but outperform the rest of two scenarios.
* The best performing algorithm seems to be the Random Forest.

Overall it seems that

* With the application of SMOT oversampling the performance improves in comparison to the strategy of without addressing the class imbalancing issue.
* The Random forest outperforms all the remaining algorithms.

**Bankruptcy data**

The bankruptcy data has 6819 rows and 96 columns. The data does not have the mission values.

I have used two criterions to remove the irrelevant columns.

1. The columns for which the correlation is high than 0.7, one of those two variables is removed.
2. The columns in which there are only 2 to 5 different values (numeric columns) means that variance is very low, are removed from the data.

For the rows which has more than 10 missing values, are removed from the data. Finally, I have 67 columns and 6819 rows. The outliers, which are away by more than 2\*SD from the mean are winsorized at the 95% level.

For most of the variables the data does not seem normal also evident in the box-plot and histogram. The t-stats for the equality of mean suggest that the data does not have same mean and most of the columns are linearly dependent as suggested by chi-square test.

Standard Transformation is applied to data to stabilized the variance and make the data ready for statistical analysis.

Machine Learning algorithms are applied in two stages.

1. In the first sage the issue of class imbalance is not addressed. Eleven different machine learning algorithms are applied and four different scenarios are considered
   1. without focusing the cross validation (CV) and features selection.
   2. Only cross validation (CV: Stratified K-Fold) is considered
   3. Only feature selection (random forest based algorithm is used) criterion is considered
   4. Both cross validation (CV: Stratified K-Fold) and features selection (random forest based algorithm is used) are considered.

The results of this stage indicate that

* With the application of CV the precision and accuracy increase in comparison to the bench mark category which is (without CV and RFFS)
* The performance in terms all four criterions (precision, recall, AUC and accuracy) worsens with the application of features selection criterion in comparison to the benchmark category
* The final Selected Features are [" Net Income to Stockholder's Equity", ' Degree of Financial Leverage (DFL)', ' Equity to Liability', ' Borrowing dependency', ' Working Capital/Equity', ' Net Value Per Share (B)', ' Total expense/Assets', ' Debt ratio %', ' ROA(C) before interest and depreciation before interest']
* With the application of both CV and RFFS criterions, almost all 11 algorithms outperform the benchmark as well as the other two categories.
* The best performing algorithm seems to be the Random Forest.

1. In the second stage the class imbalance issue is resolved through SMOT oversampling (to increase the under sampled category which is Bankruptcy=1). The results show that

Original data shape: (6819, 67)

Resampled data shape: (13198, 67)

The above four different scenarios are considered

1. without focusing the cross validation (CV) and features selection.
2. Only cross validation (CV: Stratified K-Fold) is considered
3. Only feature selection (random forest based algorithm is used) criterion is considered
4. Both cross validation (CV: Stratified K-Fold) and features selection (random forest based algorithm is used) are considered.

The results of this stage indicate that

* With the application of CV the precision and accuracy increase in comparison to the bench mark category which is (without CV and RFFS)
* The performance in terms all four criterions (precision, recall, AUC and accuracy) worsens a little bit with the application of features selection criterion in comparison to the benchmark category
* Selected Features are [" Net Income to Stockholder's Equity", ' Degree of Financial Leverage (DFL)', ' Equity to Liability', ' Borrowing dependency', ' Working Capital/Equity', ' Net Value Per Share (B)', ' Total expense/Assets', ' Debt ratio %', ' ROA(C) before interest and depreciation before interest']
* With the application of both CV and RFFS criterions, almost all 11 algorithms underperform the benchmark as well as the other two categories.
* The best performing algorithm seems to be the Random Forest followed by KNN.

Overall it seems that

* With the application of SMOT oversampling the performance improves in comparison to the strategy of without addressing the class imbalancing issue.
* The Random forest outperforms all the remaining algorithms.

**Churn Data**

The churn dataset consists of rows= 5986 and columns= 22. The cleaning function in template is called to change columns to numeric, which are ["SeniorCitizen","tenure","MonthlyCharges","TotalCharges"]. The Total Charges column has missing values which are filled by linear interpolation and un-necessary columns are deleted.

* The heatmap of correlation shows that the correlation among features is not so high.
* t-stats for equality of mean of all numeric columns suggest that mean of numeric columns are not equal
* The data of numeric columns is not normal.

The one hot encoding is applied to string columns to create dummies, except the churn column which results in

Rows=5986, and columns= 45)

Label encoding is applied to churn column. **template.apply\_label\_encoding(churndf, cols=['Churn'])**

Machine Learning algorithms are applied in two stages.

1. In the first sage the issue of class imbalance is not addressed. Eleven different machine learning classification algorithms are applied and four different scenarios are considered
   1. without focusing the cross validation (CV) and features selection.
   2. Only cross validation (CV: Stratified K-Fold) is considered
   3. Only feature selection (random forest based algorithm is used) criterion is considered
   4. Both cross validation (CV: Stratified K-Fold) and features selection (random forest based algorithm is used) are considered.

The results of this stage indicate that

* With the application of CV the precision and accuracy increase in comparison to the bench mark category which is (without CV and RFFS)
* The performance in terms all four criterions (precision, recall, AUC and accuracy) worsens with the application of features selection criterion in comparison to the benchmark category
* The final Selected Features are =['tenure', 'MonthlyCharges', 'Contract\_Month-to-month', 'Contract\_Two year', 'OnlineSecurity\_No']
* With the application of both CV and RFFS criterions, almost all 11 algorithms outperform the benchmark and RFFS based strategies however, the results are not better than the only CV based strategy.
* The best performing algorithm seems to be the XGBoost and Adaboost.

1. In the second stage the class imbalance issue is resolved through SMOT oversampling (to increase the under sampled category which is stroke=1). The results show that

Original data shape: (5986, 45)

Resampled data shape: (8798, 45)

The above four different scenarios are considered

1. without focusing the cross validation (CV) and features selection.
2. Only cross validation (CV: Stratified K-Fold) is considered
3. Only feature selection (random forest based algorithm is used) criterion is considered
4. Both cross validation (CV: Stratified K-Fold) and features selection (random forest based algorithm is used) are considered.

The results of this stage indicate that

* With the application of CV the precision and accuracy increase in comparison to the bench mark category which is (without CV and RFFS)
* The performance in terms all four criterions (precision, recall, AUC and accuracy) worsens a little bit with the application of features selection criterion in comparison to the benchmark category
* Selected Features =['tenure', 'MonthlyCharges', 'Contract\_Two year', 'Contract\_One year', 'TechSupport\_Yes', 'Contract\_Month-to-month', 'PaperlessBilling\_No']
* With the application of both CV and RFFS criterions, almost all 11 algorithms underperform than the CV only based scenario but outperform the rest of two scenarios.
* The best performing algorithm seems to be the Random Forest.

Overall it seems that

* With the application of SMOT oversampling the performance improves in comparison to the strategy of without addressing the class imbalancing issue.
* The Random forest outperforms all the remaining algorithms.