BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

Identify Fraudulent Credit Card Users

Group Number 200 - VINOD BHASKAR MAHAJAN (2019HC04139), BIRE PAYAL PRAMODRAO (2019HC04899), SURAJ KUMAR (2019HC04912)

01-Feb-2021

Introduction to Data Science

Overview

- Objective: **Identify Fraudulent Credit Card Users**
- Methodology: **Data Science Methodology based on CRISP-DM**

Dataset

- How many features: **31**
- Size of the dataset: **143MB (284,807 rows)**
- Multiple files: **No**
- What kind of data – numerical or character: **Numerical**
- Balanced or imbalanced – what is the distribution: **Imbalanced (99.83% No Frauds, 0.17% Frauds)**
- Distribution of Training set, validation set, testing set: **Training 80%, Testing 20%, Cross Validation: 5 Fold**
- Missing data and Preprocessing challenges: No Null Data, Duplicated Removed, Outliers removed from -ve correlated class, High Features Imbalance class, Scaling of Time and Amount

Feature Engineering Techniques

- Features removed: **V3, V9, V10, V12, V14, V16 and V17 are negatively correlated, V4, and V11 are positively correlated. Further applying variance inflation factor, we removed following dependent attributes V7, 17, V16, V3, V12, V10, V14, V5**
- Feature creation: **As per question, time_category, values for Amount, Time are scaled.**
- Feature ranking: (post removal) **V4, V11, V9, V2, V18, V1, V6, V19, V20, V24, V21, V27, Time, V8, V22, Amount, V13, V28, V23, V26, V15, V25**
- Class imbalance treatment: **Random Under-Sampling, removing data to have a more balanced dataset and Over Sampling (SMOTE) i.e. adding synthetic data for match class. We found that performing sampling during cross validation gives better results.**
- Any other: **Removing of outliers, duplicates. For Time attributed implemented Categorical attribute (time_category) as asked in problem statement, but haven't used this during Machine Learning**

Methodology

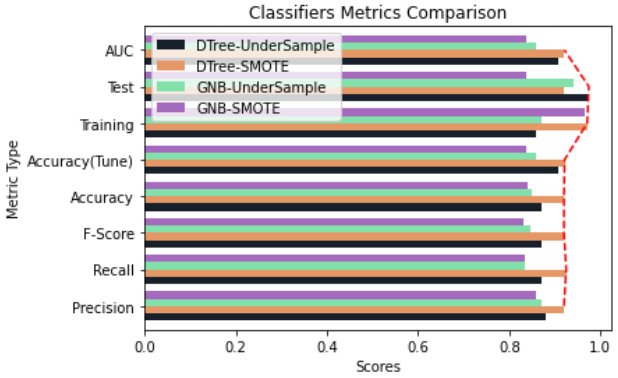
- The 2 classifiers used **(as per given problem) : Decision Tree and Naïve Based Classifier (Gaussian)**
- Ensemble pipeline: For Decision Tree imbalanced_make_pipeline for undersampling, over sampling during cross validation.
- Other models considered: We explored other implementation of Naïve Bayes classifier Multinomial, Bernoulli but finally selected Gaussian as we had most of the attributes as numerical while Multinomial is well suited for discrete while Bernoulli works on non-negative numbers and our dataset has negative values.
- Hyper-parameter tuning: We used GridSearchCV for both classifiers to the respective best estimators.
    - For Decision Tree we got (max_depth=3, min_samples_leaf=5)
    - For Gaussian Naïve Bayes we got (var_smoothing=3.511191734215127e-05)

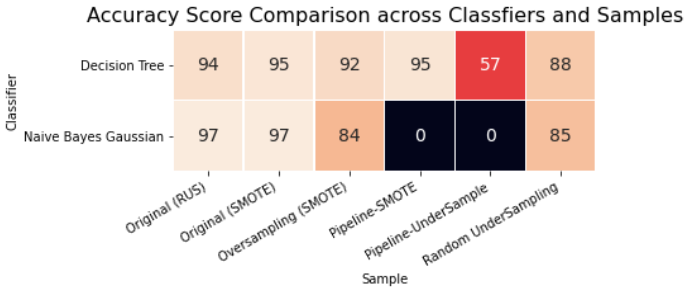
Results

- Table for the evaluation metric for each ML technique used

	Decision Tree	Naïve based Classifier (Gaussian)
Cross Validation Score (Train) accuracy	92.14%	83.66%
Accuracy	921.11%	83.61%
ROC AUC Score	0.9214	0.8368

- Plot of the curves





- Conclusion

It is observed that data is imbalanced, we handle the same we used Random under sampling and Over Sampling (SMOTE) method. We experimented and found that if sampling is done during cross validation is yields more accurate metrics as compared to overfitting and underfitting is observed. Another point is that we should store the testing data before we apply sampling method to give scores on actual data rather than synthetic data. Conclusion is that in case of imbalance class, oversampling method is better, and right way is to do that during cross validation.