Machine learning Report

Session 1: Exploratory Analysis of Imbalanced Data

For the Credit Fraud dataset, the variables that appeared strongly related to fraud behavior were: V2, V4, V10, V11, V12, V14, V16, V17, and V18.

For the Bank Marketing dataset, the variables that appeared strongly related to churn behavior were: pdays, emp.var.rate, cons.price.idx, euribor3m, nr.employed, duration, and previous.

For the Employee Attrition dataset, the variables that appeared strongly related to churn behavior were: Age, TotalWorkingYears, YearsAtCompany, YearsWithCurrManager, and EnvironmentSatisfaction.

Session 2: Churn Prediction (Part I)

For the Credit Fraud dataset, the most effective model found through grid search was: LogisticRegression(C=0.001, penalty='l2', solver='newton-cg') with an AUC of 0.98.

For the Bank Marketing dataset, the best model found during grid search was: LogisticRegression(penalty='l2', C=1e-05, solver='newton-cholesky') with an AUC of 0.93.

For the Employee Attrition dataset, the best model found through grid search was: DecisionTreeClassifier(criterion='log_loss', max_leaf_nodes=None, min_samples_split=2) with an AUC of 0.98.

Session 3: Churn Prediction (Part II)

For the Credit Fraud dataset, the best model was: LogisticRegression(C=0.001, penalty='l2', solver='newton-cg') with an AUC of 0.98, and the most influential features were: V14, V4, V12, V10, and V11.

For the Bank Marketing dataset, the best model was: RandomForestClassifier(criterion='log_loss', max_leaf_nodes=None, min_samples_split=3, n_estimators=20) with an AUC of 0.94, and the most influential features were: duration, euribor3m, age, and nr.employed.

For the Employee Attrition dataset, the best model was: RandomForestClassifier(criterion='gini', max_leaf_nodes=None, min_samples_split=2, n_estimators=50) with an AUC very close to 1, and the most influential features were: Age, MaritalStatus, StockOptionLevel, and TrainingTimesLastYear.

Session 4: Oversampling and Undersampling

Oversampling and undersampling were performed on the three datasets using SMOTE and Tomek links. Resampling improved the performance of our models.

For the Credit Fraud dataset, the best model found was:
RandomForestClassifier(criterion='log_loss', max_leaf_nodes=None, min_samples_split=2, n_estimators=100) with an AUC very close to 1.

For the Bank Marketing dataset, the best model found was:
LogisticRegression(penalty='l2', C=10, solver='liblinear') with an AUC of 0.92.

For the Employee Attrition dataset, the best model found was:
DecisionTreeClassifier(criterion='entropy', max_leaf_nodes=None, min_samples_split=2) with an AUC of 0.98.

Additional Notes:

- The use of oversampling and undersampling improves the performance of our models considerably. However, sometimes it gives strange results. The AUC found on our test sets is sometimes much higher than the AUC found through grid search, which is abnormal since thee grid search identifies the best hyperparameters based on the best AUC (obtained by averaging the AUC across test sets). Thus, it seems odd to achieve a much higher AUC with our test set.
- Random Forest models seem to consistently deliver good or even excellent results across the three different datasets, even without oversampling and undersampling. This type of model appears to perform better than others.