Jessica FORNETTI (AMSD) 17/04/2023

Rapport TPs Machine learning

Séance 1 : Analyse exploratoire des données déséquilibrées

Pour le dataset Credit Fraud, on avait trouvé que les variables qui semblent fortement liées au comportement de fraude étaient : V2, V4, V10, V11, V12, V14, V16 V17 et V18.

Pour le dataset Bank Marketing, on avait trouvé que les variables qui sembent fortement liées au comportement de churn étaient : pdays, emp.var.rate, cons.price.idx, euribor3m, nr.employed, duration et previous.

Pour le dataset Employee Attrition, on avait trouvé que les variables qui sembent fortement liées au comportement de churn étaient : Age, TotalWorkingYears, YearsAtCompany, YearsWithCurrManager et EnvironmentSatifaction.

Séance 2 : Prédiction de churn (Partie I)

Pour le dataset Credit Fraud le modèle le plus performant trouvé par le grid search est le modèle LogisticRegression(C = 0.001, penalty = 'l2', solver = 'newton-cg') avec un AUC de 0.98.

Pour le dataset Bank Marketing le meilleur modèle trouvé lors du grid search est le modèle LogisticRegression(penalty='l2', C= 1e-05, solver ='newton-cholesky') avec un AUC de 0.93

Pour le dataset Employee Attrition le meilleur modèle trouvé par le grid search est le DecsionTreeClassifier(criterion = 'log_loss', max_leaf_nodes = None, min_samples_split = 2) avec un AUC de 0.98.

Séance 3 : Prédiction de churn (Partie II)

Pour le dataset Credit Fraud, le meilleur modèle est LogisticRegression(C = 0.001, penalty = 'l2', solver = 'newton-cg') avec un AUC de 0.98 et V14, V4, V12, V10 et V11 comme les features les plus influents.

Pour le datset Bank Marketing, le meilleur modèle est RandomForestClassifier(criterion= 'log_loss', max_leaf_nodes= None, min_samples_split= 3, n_estimators= 20) avec un AUC de 0.94 et duration, euribor3m, age et nr.employed comme les features les plus influents.

Pour le datset Employee Attrition, le meilleur modèle est RandomForestClassifier(criterion = 'gini', max_leaf_nodes = None, min_samples_split = 2, n_estimators = 50) avec un AUC tres proche de 1 et Age, MaritalStatus, StockOptionLevel et TrainingTimesLastYear comme les features les plus influents.

Séance 4 : Sur-échantillonnage et sous-échantillonnage

On a effectué un oversampling puis undersampling sur les trois datasets en utilisant la méthode SMOTE et Tomek links. Le resampling permet d'améliorer la performance de nos modèles.

Pour le dataset Credit Fraud ,le meilleur modèle trouvé est le RandomForestClassifier(criterion = 'log_loss', max_leaf_nodes = None, min_samples_split = 2, n_estimators = 100) ayant un AUC très proche de 1.

Pour le dataset Bank Marketing, le meilleur modèle trouvé est la LogisticRegression(penalty='l2', C=10, solver='liblinear') ayant un AUC de 0.92.

Pour le dataset Employee Attrition, le meilleur modèle trouvé est le DecisionTreeClassifier(criterion = 'entropy', max_leaf_nodes = None, min_samples_split = 2) ayant un AUC de 0.98.

Autre remarques :

- L'utilisation de oversampling et undersampling améliore de manière considérable la performance de nos modèles. Mais parfois cela donne des résultats étranges. L'AUC trouvé sur nos test sets est parfois beaucoup plus élevé que l'AUC trouvé par le grid search ce qui est anormal car le grid search trouve les meilleurs hyperparamètres en se basant sur le meilleur AUC (qui est obtenu en faisant la moyenne de l'AUC trouvé sur chaque test set). Donc cela semble bizarre que avec notre test set on obtient un AUC beaucoup plus élevé.
- Les modèles Random Forest semblent donner de bon voir très bon résultats sur nos trois datasets différents, même sans oversampling et undersampling. Ce type de modèle a l'air plus performant que les autres modèles.