Credit Card fraud transaction detection

Apostolos Vakalos

September 2021

Εισαγωγή

Η παρούσα εργασία βασίζεται σε έναν παλιό διαγωνισμό του Kaggle με θέμα την ανίχνευση των συναλλαγών που είναι απάτες σε ένα σύνολο δεδομένων συναλλαγών. Πρόκειται για ένα πρόβλημα classification με δύο κλάσεις στο οποίο η φύση των δεδομένων είναι η ασυμμετρία των δύο κλάσεων, δηλαδή σε ένα μεγάλο πλήθος από συναλλαγές, μόνο είναι μικρό υποσύνολο είναι απάτες. Αυτή είναι και η μεγαλύτερη πρόκληση της συγκεκριμένης εργασίας.

Επεξήγηση Dataset

Time: Ο χρόνος που πέρασε σε δευτερόλεπτα από την πρώτη συναλλαγή του dataset

V1..V28: Κανονικοποιημένα (έπειτα από PCA) ανώνυμα χαρακτηριστικά συναλλαγών.

Amount: Το ποσό της συναλλαγής

Class: Η κατηγορία της συναλλαγής ως απάτη ή μη.

Προ-επεξεργασία

Η προεπεξεργασία των δεδομένων είναι σχετικά απλή με την κανονικοποίηση των χαρακτηριστικών Time και Amount με τη βοήθεια του StandarScaler εφόσον και τα χαρακτηρηστικά V1…V28 έχουν μέση τιμή στο 0.

Το επόμενο βήμα είναι η εξισορρόπηση του dataset (ολικώς ή μερικώς) με στόχο την αποφυγή του overfitting στους classifiers λόγω της συντριπτικά πλειοψηφούσας κατηγορίας των νόμιμων συναλλαγών. Θα βοηθήσουμε τους classifiers να μην αγνοούν τα σχετικά λίγα παραδείγματα συναλλαγών-απατών και να μάθουν να ανιχνεύουν τις απάτες δίνοντας τους κατά τη διάρκεια της εκπαίδευσης (σχεδόν) ίσα παραδείγματα από κάθε κατηγορία.

Επομένως η εξισορρόπηση των κατηγοριών θα γίνει μόνο στο train set ενώ το test set θα παραμείνει ανέπαφο.

Η προσπάθεια εξισορρόπησης εστίασε κυρίως στη σύγκριση και παραμετροποίηση των διαφόρων μεθόδων και χωρίζεται σε δύο κατηγορίες.

- Undersampling στις νόμιμες συναλλαγές:

    - Random Undersampling
    - Near Miss
    - Condensed Nearest Neighbors - CNN
    - Edited Nearest Neighbors - ENN
    - Tomek Links

- One Sided Selection (CNN + Tomek Links)
- Neighborhood Cleaning Rule (CNN + ENN)

➢ Undersampling και Oversampling στις νόμιμες και παράνομες συναλλαγές.

- SMOTETomek
- SMOTEENN

Η πλήρης εξισορόπηση είχε σαν αποτέλεσμα πολύ μικρό train set και χαμηλότερη αποδοση από το αρχικό train set. Για αυτό το λόγο προτιμήθηκαν κυρίως οι OSS και NCR με την τελευταία να είναι και η τελική επιλογή λόγω καλύτερης απόδοσης.

Αλγόριθμοι Ταξινόμησης

Έχοντας ένα πιο αντιπροσωπευτικό για την κατηγορία fraud, train set δοκιμάσαμε διάφορους αλγορίθμους ταξινόμησης και τους αξιολογήσαμε ως προς το precission, recall και το AUPRC (Area Under Precision-Recall Curve). Το accuracy στη συγκεκριμένη περίπτωση δεν αποτελεί αξιόπιστο μέτρο αξιολόγησης.

Οι μέθοδοι που δοκιμάσαμε είναι οι παρακάτω.

- Logistic Regression
- SVM
- kNN
- XGBoost
- MLP

Θα συγκρίνουμε την απόδοση τους στο resampled και στο αρχικό σύνολο εκπαίδευσης για να δούμε τις διαφορές.

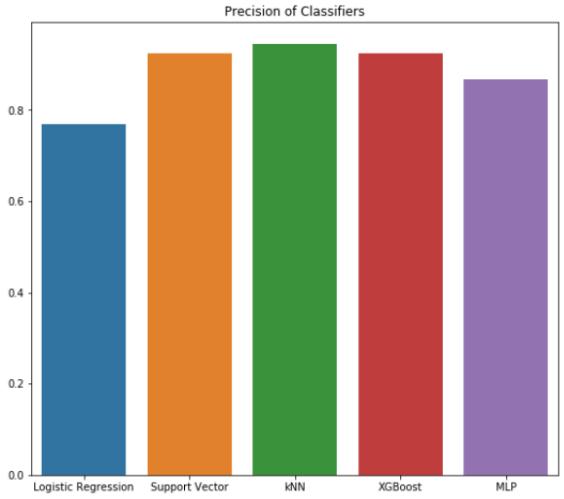
Θα δούμε πρώτα το initial train set και μετά το resampled.

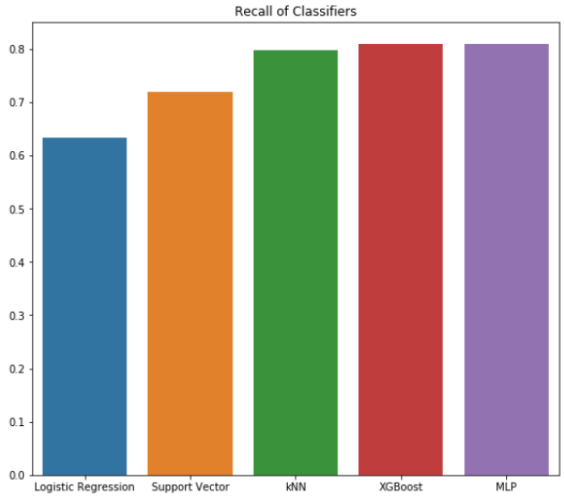
Initial train set

```
------------------------------------------------------

XGBoost Classifier

------------------------------------------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     93834
           1       0.93      0.81      0.86       153

    accuracy                           1.00     93987
   macro avg       0.96      0.91      0.93     93987
weighted avg       1.00      1.00      1.00     93987

True Positives: 124
False Negatives: 29
True Negatives: 93824
False Positives: 10
```

```
------------------------------------------------------

kNN Classifier

------------------------------------------------------
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     93834
           1       0.95      0.80      0.87       153

    accuracy                           1.00     93987
   macro avg       0.97      0.90      0.93     93987
weighted avg       1.00      1.00      1.00     93987

True Positives: 122
False Negatives: 31
True Negatives: 93827
False Positives: 7
```

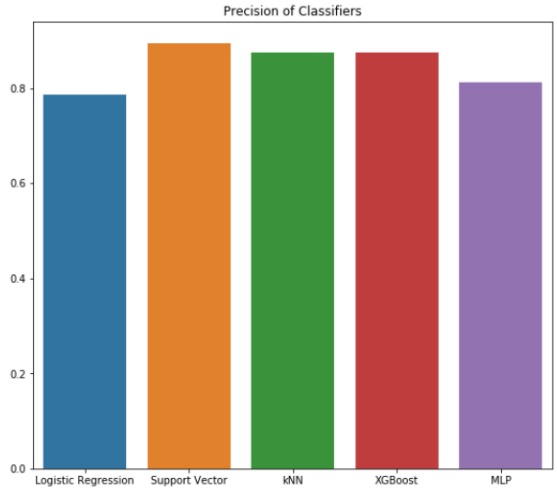
Παρά την ασυμμετρία μεταξύ των κλάσεων οι ταξινομητές πετυχαίνουν αρκετά καλό precision και recall με τις καλύτερες μεθόδους να είναι οι kNN και XGBoost.

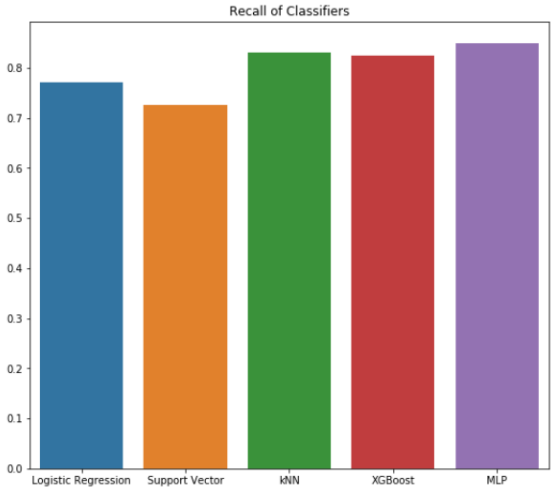




Resampled train set

Το καλύτερο μοντέλο είναι ο kNN με precision 95% και recall 80%.





Η απόδοση είναι παραπλήσια με καλύτερο recall όμως λόγω του πιο ισορροπημένου dataset με αποτέλεσμα να αναγνωρίζονται καλύτερα τα παραδείγματα της κλάσης fraud. Το precision φαίνεται ελαφρώς μειωμένο. Το καλύτερο μοντέλο είναι ο kNN με precision 88% και recall 83%.

Συμπερασματικά, δεν παρατηρήθηκε η αναμενόμενη βελτίωση στα μοντέλα με το resampling. Αυτό μας οδηγεί στη διαπίστωση ότι ίσως τα παραδείγματα των δύο κλάσεων να είναι αρκετά διαχωρισμένα (high separation) ώστε ακόμα και χωρίς κάποια ιδιαίτερη εξισορρόπηση να μπορούν να αναγνωριστούν από τους classifiers.

K-Fold Cross Validation

Μια εναλλακτική προσέγγιση ήταν του 5-fold Cross Validation. Χρησιμοποιήθηκε η μέθοδος stratified k fold με suffling των δεδομένων. Κατόπιν εφαρμόζεται ένας αλγόριθμος δειγματοληψίας των δεδομένων ώστε να διορθώσει την ασυμετρία των κλάσεων και μετά εφαρμόζεται μια μέθοδος ταξινόμησης. Με αυτή την τεχνική δοκιμάστηκαν τα δύο καλύτερα μοντέλα το KNN και το XGBoost.

Παρατηρήσεις:

Η ταχύτητα του αλγορίθμου δειγματοληψίας είναι κομβικής σημασίας. Πλέον τα train sets είναι πολύ μεγάλα και οι πιο σύνθετες μέθοδοι δειγματοληψίας καθυστερούν υπερβολικά. Η μέθοδος SMOTE πέτυχε τα καλύτερα αποτελέσματα και σε εύλογο χρόνο.

Εγινε σύγκριση με το δειγματοληπτημένο σύνολο και με το αρχικό. Τα ευρήματα επιβεβαιώνουν το αρχικό συμπέρασμα, ότι μετά τη δειγμτοληψία έχουμε καλύτερο recall αλλά χειρότερο precision όπως φαίνεται στην εικόνα:

XGBoost με SMOTE:

```
------------------------------------------------------------
Total prefromance for 5-Fold Cross Validation

Average precision: 0.7632425638484386  Average Recall: 0.8373118944547515
```

XGBoost χωρίς δειγματοληψία:

```
------------------------------------------------------------
Total prefromance for 5-Fold Cross Validation

Average precision: 0.9572848934603894  Average Recall: 0.804803133374562
```

KNN με SMOTE:

```
------------------------------------------------------------
Total prefromance for 5-Fold Cross Validation

Average precision: 0.5875983427923938  Average Recall: 0.8332714904143476
```

KNN χωρις δειγματοληψία:

```
------------------------------------------------------------
Total prefromance for 5-Fold Cross Validation

Average precision: 0.9278791617106978  Average Recall: 0.7864564007421151
```

Τελικά παρατηρείται μία υπεροχή του XGBoost έναντι του KNN και στις δύο μετρικές.