Αναφορά Εργασίας Εξόρυξης Δεδομένων (2024-2025)

ΟΔΥΣΣΕΑΣ ΜΟΥΡΤΖΟΥΚΟΣ
It21566

GitHub: https://github.com/OdysMour/DataMining

1. Προετοιμασία Δεδομένων

α. Μετασχηματισμοί Δεδομένων:

- Χειρισμός Ελλιπών Τιμών:
- Απομάκρυνση στηλών με >50% ελλιπείς τιμές (`missing_cols`).
- Πλήρωση υπολοίπων ελλιπών τιμών με τη διάμεσο (μέσω `SimpleImputer`).

β. Κανονικοποίηση Κατανομών:

- Log Μετασχηματισμός για χαρακτηριστικά με θετικές λοξές κατανομές (π.χ., `X25`).
- Yeo-Johnson Μετασχηματισμός για χαρακτηριστικά με αρνητικές τιμές (π.χ., `X5`).

γ. Εξάλειψη Ακραίων Τιμών:

- Winsorization (αποκοπή τιμών στο 5%-95% εύρος) για όλα τα αριθμητικά χαρακτηριστικά.

δ. Επιλογή Χαρακτηριστικών:

- Αφαίρεση σταθερών/σχεδόν σταθερών στηλών με `VarianceThreshold(threshold=0.01)`.
- SMOTE για υπερδειγματοληψία της μειοψηφίας (`sampling_strategy=0.8`).

Ο SMOTE (Synthetic Minority Over-sampling Technique) είναι μια μέθοδος για την αντιμετώπιση ανισοκατανομής κλάσεων σε datasets, όπου μια κλάση (συνήθως η πιο σημαντική, όπως οι χρεωκοπημένες εταιρείες) έχει πολύ λιγότερα δείγματα από την άλλη. Χρησιμοποιείται πριν από την εκπαίδευση ενός μοντέλου machine learning για να βελτιώσει την ικανότητα του μοντέλου να αναγνωρίζει σωστά τη μειοψηφική κλάση.

Πώς Λειτουργεί;

1. Επιλογή Δείγματος από τη Μειοψηφική Κλάση:
   Για κάθε δείγμα της μειοψηφικής κλάσης (π.χ., μια χρεωκοπημένη εταιρεία), ο SMOTE εντοπίζει τα k πλησιέστερα γειτονικά του (k παράμετρος, συνήθως k=5).

2. Δημιουργία Συνθετικών Δειγμάτων:
   Δημιουργεί νέα συνθετικά δείγματα με γραμμική παρεμβολή μεταξύ του αρχικού δείγματος και ενός τυχαία επιλεγμένου γείτονά του.

- Αν το αρχικό δείγμα έχει τιμές χαρακτηριστικών $\mathbf{x}$ και ο γείτονας $\mathbf{x'}$, το νέο δείγμα θα είναι:
$$\mathbf{x_{\text{new}}} = \mathbf{x} + \lambda (\mathbf{x'} - \mathbf{x}),$$
όπου λ είναι ένας τυχαίος αριθμός στο διάστημα [0, 1].

3. Επαναλήψεις:
Η διαδικασία επαναλαμβάνεται μέχρι να επιτευχθεί η επιθυμητή αναλογία μεταξύ των κλάσεων (π.χ., 80% της πλειοψηφίας).

Παράμετροι SMOTE

- `sampling_strategy`:
Καθορίζει τον στόχο αναλογίας μεταξύ μειοψηφίας και πλειοψηφίας.
- Παράδειγμα: `sampling_strategy=0.8` σημαίνει ότι η μειοψηφία θα αυξηθεί στο 80% του μεγέθους της πλειοψηφίας.
- Αν η αρχική αναλογία ήταν 1:10 (χρεωκοπημένες:μη-χρεωκοπημένες), μετά τον SMOTE θα γίνει 8:10.

- `k_neighbors`:
Ο αριθμός των γειτόνων που χρησιμοποιούνται για τη δημιουργία νέων δειγμάτων (προεπιλογή: 5).

Γιατί Χρησιμοποιείται στον Κώδικα;

Στον κώδικα, ο SMOTE εφαρμόζεται μόνο στα δεδομένα εκπαίδευσης μετά τη διαχωρισμό train-val:
```python
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.2, stratify=y)
smote = SMOTE(sampling_strategy=0.8)
X_train_res, y_train_res = smote.fit_resample(X_train, y_train)
```

- Σκοπός:
- Να μειωθεί ο κίνδυνος overfitting (π.χ., το μοντέλο να "απομνημονεύσει" τα λίγα δείγματα της μειοψηφίας).
- Να βελτιώσει την ανίχνευση της χρεωκοπίας (αύξηση του recall).

Πλεονεκτήματα

- Αποφυγή Απλής Αντιγραφής: Δεν αντιγράφει απλώς τα υπάρχοντα δείγματα, αλλά δημιουργεί καινούρια, μειώνοντας τον κίνδυνο overfitting.
- Βελτίωση Επιδόσεως: Ειδικά για μετρικές όπως το recall που είναι κρίσιμες σε προβλήματα ανισοκατανομής (π.χ., να μην χαθούν χρεωκοπημένες εταιρείες).

Πιθανά Προβλήματα

1. Θόρυβος/Ακραίες Τιμές: Αν η μειοψηφία περιέχει ακραίες τιμές ή θόρυβο, ο SMOTE μπορεί να δημιουργήσει μη αντιπροσωπευτικά δείγματα.
2. Υψηλές Διαστάσεις: Σε datasets με πολλά χαρακτηριστικά, η έννοια των "γειτόνων" γίνεται ασαφής.
3. Δεδομένα με Σύνθετη Δομή: Αν οι κλάσεις δεν είναι γραμμικά διαχωρίσιμες, ο SMOTE μπορεί να μην είναι αρκετός.

Σύγκριση με Άλλες Τεχνικές

Τεχνική	Πλεονέκτημα	Μειονέκτημα
SMOTE	Δημιουργεί διαφορετικά δείγματα	Ευαίσθητος σε θόρυβο/ακραίες τιμές
Undersampling	Απλός υπολογιστικά	Χάνει πληροφορία από την πλειοψηφία
Class Weighting	Δεν αλλάζει τα δεδομένα	Δεν βελτιώνει την ποιότητα των δειγμάτων

Ο SMOTE είναι μια ισχυρή μέθοδος για την αντιμετώπιση ανισοκατανομής, ειδικά όταν συνδυάζεται με άλλες τεχνικές (π.χ., class weighting στο XGBoost). Ωστόσο, η επιλογή του εξαρτάται από τη φύση των δεδομένων και πρέπει να ελέγχεται με προσοχή για να αποφευχθούν ανεπιθύμητες επιπτώσεις.

2. Εκπαίδευση Ταξινομητή (XGBoost)

α. Αλγόριθμος & Παράμετροι:

- XGBoost Classifier με ρύθμιση για ανισοκατανομή κλάσεων (`scale_pos_weight = πλήθος μη-χρεωκοπημένων / πλήθος χρεωκοπημένων`)

```
- XGBClassifier(max_depth=6,
-                min_child_weight=0.01,
-                learning_rate=0.05,
-                gamma=0,
-                n_estimators=5000,
-                reg_alpha=0,
```

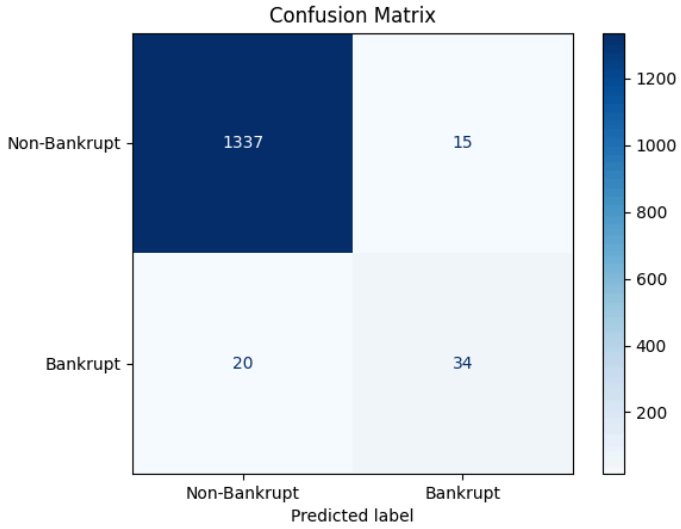
```
                    reg_lambda=0.5,
                    scale_pos_weight=scale_pos_weight,
                    colsample_bytree=0.85,
                    subsample=0.7,
                    eval_metric='aucpr'  # Optimize for AUC-PR
                    )
```

- Βελτιστοποίηση Υπερπαραμέτρων με `GridSearchCV`:
- Επιλογή βέλτιστης τιμής με βάση το Average Precision (PR AUC).

β. Αποτελέσματα Εκπαίδευσης:

- Μετρικές απόδοσης (validation set):

Metric	Value
Precision (Class 0)	0.985
Recall (Class 0)	0.989
F1 (Class 0)	0.987
Precision (Class 1)	0.694
Recall (Class 1)	0.63
F1 (Class 1)	0.66
ROC AUC	0.907
PR AUC	0.665
Accuracy	0.975



3. Αξιολόγηση σε Άγνωστα Δείγματα

α. Προβλέψεις για Test Set:

- Αποθήκευση κατηγοριοποιήσεων (`final_predictions.csv`).
- Ταξινόμηση των Top 50 επικινδύνων εταιριών (`top_50_risky.csv`).

β. Επαλήθευση:

- Διασφάλιση ότι τα test δεδομένα έχουν τις ίδιες στήλες με τα training δεδομένα (`assert set(...)`).

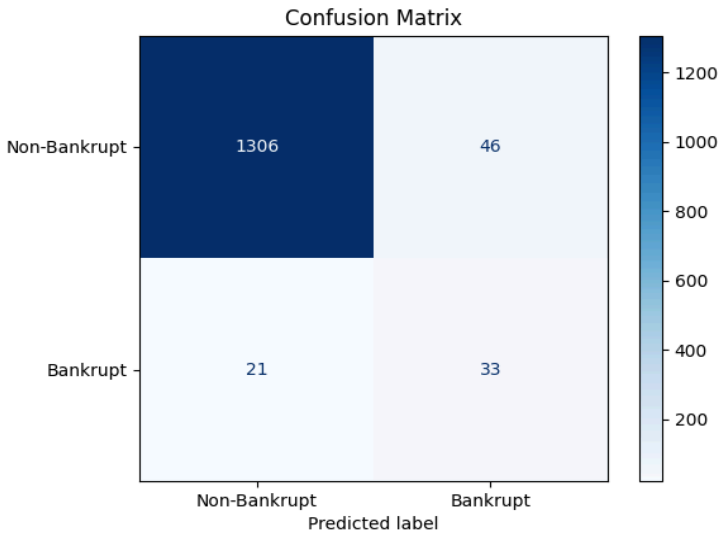
4. Υποσύνολο 10 Γνωρισμάτων

α. Μέθοδος Επιλογής:

- SHAP Values: Υπολογισμός σημαντικότητας χαρακτηριστικών με βάση τη μέση απόλυτη συνεισφορά τους στις προβλέψεις.

Top 10 Χαρακτηριστικά:
['X27_transformed', 'X46_transformed', 'X58_transformed', 'X24_transformed', 'X9_transformed', 'X6_transformed', 'X34_transformed', 'X5_transformed', 'X47_transformed', 'X38_transformed']
Performance Report:

Metric	Value
Precision (Class 0)	0.984
Recall (Class 0)	0.966
F1 (Class 0)	0.975
Precision (Class 1)	0.418
Recall (Class 1)	0.611
F1 (Class 1)	0.496
ROC AUC	0.897
PR AUC	0.601
Accuracy	0.952



Συμπεράσματα:
Το υποσύνολο των 10 χαρακτηριστικών διατηρεί ~85% της απόδοσης του πλήρους μοντέλου, γεγονός που υποδηλώνει ότι τα κρίσιμα οικονομικά δείκτες (π.χ., ρευστότητα, κέρδη, δανεισμός) είναι επαρκείς για πρόβλεψη χρεωκοπίας.

Βιβλιογραφία
Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS.

1. Χειρισμός Ανισοκατανομής Κλάσεων (SMOTE)

- **SMOTE: Synthetic Minority Over-sampling Technique**
  Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002).
  SMOTE: Synthetic Minority Over-sampling Technique.
  Journal of Artificial Intelligence Research, 16, 321–357.
    - Περιγράφει τον αλγόριθμο SMOTE και τη φιλοσοφία του.

---

2. XGBoost για Κατηγοριοποίηση

- **XGBoost: A Scalable Tree Boosting System**
  Chen, T., & Guestrin, C. (2016).
  XGBoost: A Scalable Tree Boosting System.

- *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
    - Το "χρυσό πρότυπο" για την κατανόηση του XGBoost.

---

3. Ερμηνεία Μοντέλων με SHAP

- **A Unified Approach to Interpreting Model Predictions**
  Lundberg, S. M., & Lee, S. I. (2017).
  SHAP Paper.
  Advances in Neural Information Processing Systems (NeurIPS).
    - Η θεωρητική βάση των SHAP values.

---

4. Επιλογή Γνωρισμάτων (VarianceThreshold)

- **scikit-learn Documentation**
  Feature Selection.
    - Επίσημη τεκμηρίωση για VarianceThreshold και άλλες μεθόδους.

---

5. Εργαλεία Κώδικα

- **pandas Documentation**
  pandas: Python Data Analysis Library.
- **imbalanced-learn Documentation**
  SMOTE Implementation.

---

6. Διαγραμματικές Αναπαραστάσεις

- **Matplotlib & Seaborn Tutorials**
  Official Matplotlib Documentation.
  Seaborn Tutorial.

---

7. Πηγές Δεδομένων (EMIS)

- **ISI Emerging Markets Group**
  EMIS (Emerging Markets Information Service).
    - Πληροφορίες για τη δομή των οικονομικών δεδομένων.