ΕΡΓΑΣΙΑ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ

Βασίλειος Κοκκινογένης it2021042

Α) Προετοιμασία

Για την προετοιμασία των δεδομένων χρειάστηκαν να γίνουν αρκετοί μετασχηματισμοί. Αρχικά, αντικατέστησα τις τιμές που είχαν ? με NaN και μετέτρεψα τις τιμές που είχαν τύπο object σε τύπο numeric, ώστε να μπορέσω να κάνω αργότερα fillna με τον μέσο όρο των υπόλοιπων τιμών. Ωστόσο πριν επιχειρήσω να κάνω fillna βρήκα τις στήλες (χαρακτηριστικά) που είχαν το μεγαλύτερο ποσοστό missing values και επιχείρησα να τις κάνω drop, μήπως βελτιωθεί η επίδοση του μοντέλου, κάτι που όμως δεν έγινε.

Β) Κατηγοριοποίηση

Για την κατηγοριοποίηση αντί για train test split με κάποιο seed χρησιμοποίησα k-fold validation για να έχω καλύτερη εικόνα για την επίδοση του μοντέλου στα άγνωστα δείγματα (σε αυτά που δεν έχει κάνει fit). Επιπλέον, επειδή το dataset είναι imbalanced (οι εταιρείες που δεν έχουν χρεωκοπήσει είναι πολύ περισσότερες σε αριθμό από αυτές που χρεωκόπησαν), έκανα SMOTE στα δεδομένα εκπαίδευσης και έπειτα κανονικοποίηση με MinMaxScaler. Όσον αφορά το μοντέλο κατηγοριοποίησης, δοκίμασα διάφορα μοντέλα (LogisticRegression, RandomForestClassifier, DecisionTreeClassifier, XGBClassifier) από τα οποία διάλεξα το XGBClassifier, μιας και είχε την καλύτερη επίδοση στην κατηγοριοποίηση στα δεδομένα εκπαίδευσης και στα δεδομένα ελέγχου. Πιο συγκεκριμένα, στα δεδομένα εκπαίδευσης είχε:

```
-------- Predictions On Training Data --------
Average Confusion Matrix:
 [[5401    4]
 [   2 5403]]
Average Accuracy: 0.9994634323703071
Average Precision: 0.9992234067697273
Average Recall: 0.9997039572564537
Average F1-Score: 0.999463574370463
```

Ενώ στα δεδομένα ελέγχου είχε:

```
-------- Predictions On Testing Data --------
Average Confusion Matrix:
 [[1333   18]
 [  20   34]]
Average Accuracy: 0.9726770374045145
Average Precision: 0.6496360453191089
Average Recall: 0.6308417508417509
Average F1-Score: 0.6378228398030357
```

Χρησιμοποίησα τις default παραμέτρους για το μοντέλο και για να αποφύγω το overfitting, χρησιμοποίησα early_stopping_rounds, το οποίο σταματάει πιο νωρίς την εκπαίδευση άμα δε βελτιώνεται πια η επίδοση στα δεδομένα ελέγχου για έναν συγκεκριμένο αριθμό rounds.

Γ) Αξιολόγηση Γνωρισμάτων – Παλινδρόμηση

Για να βρω ένα υποσύνολο 10 το πολύ γνωρισμάτων που συμμετέχουν περισσότερο στην πρόβλεψη, πήρα τα feature_importances του μοντέλου μετά την εκπαίδευση (τον μέσο όρο από όλα τα folds) και κράτησα τα 10 πρώτα features, αφότου τα ταξινόμησα πρώτα με βάση το importance τους (desc).

Κρατώντας μόνο αυτά τα 10 features και κάνοντας drop τα υπόλοιπα δεν είχαμε θετική επίδραση στην απόδοση του μοντέλου στην κατηγοριοποίηση. Πιο συγκεκριμένα, είχαμε:

```
Top 10 features with their importances and positions:
Mean Importance: ('X21', 0.12274395)
Mean Importance: ('X27', 0.1037959)
Mean Importance: ('X26', 0.05168449)
Mean Importance: ('X16', 0.04915342)
Mean Importance: ('X34', 0.049136527)
Mean Importance: ('X25', 0.03143257)
Mean Importance: ('X24', 0.029880866)
Mean Importance: ('X6', 0.029538745)
Mean Importance: ('X37', 0.026292905)
Mean Importance: ('X13', 0.02500605)
```

Και επίδοση του μοντέλου στα δεδομένα εκπαίδευσης:

```
-------- Predictions On Training Data --------
Average Confusion Matrix:
 [[5310   95]
 [  12 5393]]
Average Accuracy: 0.9901568064496029
Average Precision: 0.9827851629213781
Average Recall: 0.9978167061641392
Average F1-Score: 0.9902389334722571
```

Ενώ στα δεδομένα ελέγχου:

```
-------- Predictions On Testing Data --------
Average Confusion Matrix:
 [[1283   68]
 [  22   32]]
Average Accuracy: 0.9361047468146177
Average Precision: 0.3282619531634173
Average Recall: 0.5976430976430976
Average F1-Score: 0.42050306570520257
```

Βλέπουμε ότι η επίδοση έχει μειωθεί σημαντικά, με F1-Score κοντά στο 0.4 σε σχέση με πριν που ήταν κοντά στο 0.6

Είχα δοκιμάσει και με άλλους τρόπους να βρω τα 10 σημαντικότερα features (είτε με RFECV, που είναι RFE με k-fold validation, είτε με apriori, λαμβάνοντας ως υπόψη μόνο τους κανόνες που είχαν ως consequent το γνώρισμα X65 της χρεωκοπίας), αλλά και πάλι είχα παρόμοια αποτελέσματα με F1-Score κοντά στο 0.4