Big HW Report

Alexander Belov

December 17, 2023

1 Link to Github repo

https://github.com/fourlex/osda_hw3

2 Datasets and preprocessing

I've prepared two versions of each dataset. The first version only has categorical variables one-hot encoded; numerical vars are left as is. This version is used for FCA with pattern structures. In the second version, in addition to categorical vars, numeric variables are binarized (with bin boundaries listed in [ ] below for each variable), and 5-point scores are thresholded into 1 if the value is ≥ 4. The second versions of the datasets is used for all other classification models.

- Dataset №1: Hotel Bookings.
  - Target variable:
    - 'is_cancelled' — whether the hotel reservation was cancelled.
  - categorical:
    - 'deposit_type', 'customer_type', 'market_segment', 'meal', 'hotel'
  - numeric:
    - 'lead_time': [-0.01, 100, 200, 400, ∞],
    - 'stays_in_week_nights': [-0.01, 3, ∞],
    - 'total_of_special_requests': [-0.01, 1, ∞],
    - 'days_in_waiting_list': [-0.01, 1, ∞],
    - 'adults': [-0.01, 1, ∞],
    - 'children': [-0.01, 0, ∞] – number of children in the reservation. This column had lots of missing values I replaces with 0s.
    - 'babies': [-0.01, 0, ∞],
    - 'booking_changes': [-0.01, 0, ∞],
- Dataset №2: Heart Disease. This dataset is with low number of positives.
  - Target variable:
    - 'HeartDiseaseorAttack' — whether the person had a stroke.
  - categorical (were one-hot encoded):
    - 'Diabetes' — type of diabetes this person has
  - numeric :
    - 'BMI': [-0.01, 20, 30, 40, ∞] — body mass indicator
    - 'PhysHlth': [-0.01, 0, 10, ∞] — assesment of physical health
    - 'MentHlth': [-0.01, 0, 10, ∞] — assesment of mental health
    - 'Age': [-0.01, 5, 10, ∞] — relative age
    - 'Income': [-0.01, 3, 6, 7, ∞] — relative income

    - ∗ 'Education': [-0.01, 3, 5, ∞] — additional years of education
  - 5-point scores
    - ∗ 'GenHlth' — 5-point score assesment of general health

- Dataset №3: Airline passenger satisfaction.
  - target variable:
    - ∗ 'satisfaction' — whether the client was satisfied with the service
  - categorical:
    - ∗ 'Gender', 'Customer Type', 'Type of Travel', 'Class',
  - 5-point scores:
    - ∗ 'Inflight wifi service', 'Departure/Arrival time convenient', 'Ease of Online booking', 'Gate location', 'Food and drink', 'Online boarding', 'Seat comfort', 'Inflight entertainment', 'On-board service', 'Leg room service', 'Baggage handling', 'Checkin service', 'Inflight service', 'Cleanliness'
  - numerical:
    - ∗ 'Age': [-0.01, 20, 40, 60, ∞],
    - ∗ 'Flight Distance': [-0.01, 1000, 2000, 3000, 4000, ∞],
    - ∗ 'Departure Delay in Minutes': [-0.01, 60, 4*60, ∞],
    - ∗ 'Arrival Delay in Minutes': [-0.01, 60, 4*60, ∞],

3 Experiments

I perform grid search over the following binary classification models and hyperparameters:

```
{
'naive_bayes': (BernoulliNB(binarize=False), {
    'alpha': [1.0,]
}),
'xgboost': (XGBClassifier(random_state=0, objective='binary:logistic'), {
    'n_estimators': [100, 200],
}),
'random_forest': (RandomForestClassifier(random_state=0), {
    'n_estimators': [100, 200],
}),
'logreg': (LogisticRegression(random_state=0), {
    'C': [0.1, 1],
    'class_weight': [None, 'balanced']
}),
'knn': (KNeighborsClassifier(), {
    'n_neighbors': [5, 10],
    'weights': ['distance'],
}),
'catboost': (CatBoostClassifier(verbose=0, random_state=0), {
    'n_estimators': [100, 200],
}),
'lazy_fca': (MyBinarizedBinaryClassifier(), {
    'alpha': [0, 0.1, 0.5, 0.9, 1.0],
    'method': ['standard', 'standard-support', 'ratio-support'],
}),
'lazy_fca_pat_structures': (MyPatternBinaryClassifier(), {
    # 'alpha': [0, 0.1, 0.5, 0.9, 1.0],
    'alpha': [0.5, 0.9],
    'method': ['standard', 'standard-support', 'ratio-support'],
```

```
        # 'method': ['standard', 'standard-support', 'ratio-support'],
    })
}
```

Each dataset is divided into train set (700 samples) and test set (100 samples). I perform stratified 5-fold CV on train dataset for each model and hyperparameter. The hyperparameters with the highest mean f1 score on CV are chosen for each model. Each model is then refitted on train set with those hyperparameters.

4 Results

Refitted models are then used to predict target variables on test parts of the datasets. The resulted classification metrics are presented below.

dataset	airline		heart		hotel	
	accuracy	f1	accuracy	f1	accuracy	f1
model						
catboost	0.830	0.800	0.880	0.143	0.780	0.656
knn	0.800	0.767	**0.910**	**0.400**	0.690	0.551
lazy_fca	0.730	0.743	0.690	0.311	0.690	0.644
lazy_fca_pat_structures	**0.840**	**0.833**	0.690	0.311	0.540	0.610
logreg	0.820	0.800	0.720	0.364	0.650	0.588
naive_bayes	**0.840**	0.818	0.820	0.308	**0.820**	**0.710**
random_forest	0.830	0.800	0.890	0.154	0.780	0.667
xgboost	0.830	0.805	0.860	0.222	0.770	0.667