Purchase Suppression - Online Model

Status 12/18/2024:

- [x] Lopez Camey, Carlos Eduardo(AWF) created a balanced training data with "labelPurchasedLeafClick" labels (i.e., we consider only users that purchased an item from the leaf and label will tell if user is still interested
in the category within next 60 days of impression time.
- [x] Config to train Simplex model:

```
{
    "algo_family_name": ["RecentlyViewedItemsV2MobileWithMLRV6RankerPricelessTop50Features"],
    "calibration_type": "exponential_uniform_0.5",
    "hdfs_train_data_dir": "/apps/b_perso/ps_prod_balanced_smaller_v2/pretrainer",
    "input_date": "20240825-20240830",
    "label_column": "labelPurchasedLeafClick",
    "model_tag": "ps_v1",
    "ems_project_name": "be-simplex-model",
    "github_org": "<username>",
    "batch_user": "b_perso",
    "hdfs_models_store_dir": "/apps/b_perso/vlp/simplark/models/root"
}
```

```
from model_configs.base import ModelConfig

ps_v1_training_params = {
    'tree_method': 'hist',
    'booster': 'gbtree',
    'objective': 'binary:logistic',
    'random_state': 42,
    'learning_rate': 0.01,
    'colsample_bytree': 0.5,
    'eta': 0.3,
    'max_depth': 6,
    'n_estimators': 100,
    'subsample': 0.75,
    'alpha': 100,
}

V1_FEATURES = ["NumSameRviLeafCatInLastDay", "NumSameRviLeafCatInLastTwoDay",
"NumSameRviLeafCatInLastThreeDay", "NumSameRviLeafCatInLastWeek",
"PoissonNextEventProbSameLeafCatIdInRvi", "FreqSameLeafCatIdInRvi", "FreqSameLeafCatIdInWatch",
                    "PoissonNextEventProbSameLeafCatIdInWatch", "FreqSameLeafCatIdInTransaction",
"PoissonNextEventProbSameLeafCatIdInTransaction", "TimeDiffFromLastTransaction",
"TimeDiffFromLastTransactionFar"]
ps_v1 = ModelConfig(model_tag="ps_v1",
                        features=V1_FEATURES_LEAF_CAT_ONLY,
                        training_params=ps_v1_training_params)
```

- [x] Trained model: https://github.ebay.com/hroitman/treeplex/tree/krylov_week50_20241218045337_Wednesday/htmls/ps_v1
- [ ] Trained model:

20240825-2024083- https://github.ebay.com/olivyatan/treeplex/tree/krylov_week00_20250101132439_Wednesday/htmls/ps_v1

- [ ] Metrics:

	Train	Holdout
AUC	0.75908	0.75517
Precision	0.66823	0.66551
Recall	0.75900	0.75587

Next steps:

- Add filter in data preparation step that will remove duplicate leaf categories from each impression
  This is expected to reduce noise in model training and provide more robust metrics.
- Implement additional category-level features:
  Currently we use all existing Simplark features that are on leaf category level. They are too restrictive (on exact leaf match) and we need more flexibility (i.e., extend to L2 and meta categories and similarity-based features to estimate propensities).
  The following is a concrete list of features we need, which are further easy to implement in Simplark (yet that would require to add support for leaf names and L2+meta categories IDs and names from Cassini?)
- ["NumSameRviL2CatInLastDay", "NumSameRviL2CatInLastTwoDay", "NumSameRviL2CatInLastThreeDay", "NumSameRviL2CatInLastWeek", "PoissonNextEventProbSameL2CatIdInRvi", "FreqSameL2CatIdInRvi", "FreqSameL2CatIdInWatch", "PoissonNextEventProbSameL2CatIdInWatch", "FreqSameL2CatIdInTransaction", "PoissonNextEventProbSameL2CatIdInTransaction"]
- ["NumSameRviMetaCatInLastDay", "NumSameRviMetaCatInLastTwoDay", "NumSameRviMetaCatInLastThreeDay", "NumSameRviMetaCatInLastWeek", "PoissonNextEventProbSameMetaCatIdInRvi", "FreqSameMetaCatIdInRvi", "FreqSameMetaCatIdInWatch", "PoissonNextEventProbSameMetaCatIdInWatch", "FreqSameMetaCatIdInTransaction", "PoissonNextEventProbSameMetaCatIdInTransaction"]
- Instead of item title text similarity implement category name similarity (see reference implementations here: https://github.corp.ebay.com/pl/simplark/blob/e5ed66565e7ca9914736618d4d7b67ac4ea2efe5/simplark-features/src/main/scala/com/ebay/pl/datapipe/features/impl/watch/handlers.scala)

```
#Leaf category name similarity features
LeafNameJaccardSimilarityToRviCentroid
LeafNameCosineSimilarityToRviCentroid
LeafNameCosineSimilarityToFirstRvi
LeafNameCosineSimilarityToFirstTwoRvisCentroid
LeafNameCosineSimilarityToFirstThreeRvisCentroid
LeafNameCosineSimilarityToFirstFiveRvisCentroid
LeafNameCosineSimilarityCentroidRvisInLastDay
LeafNameCosineSimilarityCentroidRvisInLastTwoDays
LeafNameCosineSimilarityCentroidRvisInLastThreeDays
LeafNameCosineSimilarityCentroidRvisInLastFiveDays
LeafNameCosineSimilarityCentroidRvisInLastWeek
LeafNameCosineSimilarityCentroidRvisInLastHour
LeafNameCosineSimilarityCentroidRvisInLastTwoHours
LeafNameCosineSimilarityCentroidRvisInLastFiveHours
LeafNameCosineSimilarityCentroidRvisInLastHalfDay
LeafNameCosineSimilarityWeightedCentroidRvisRankDist
LeafNameCosineSimilarityWeightedCentroidRvisTimeDecay
LeafNameJaccardSimilarityToSearchQueriesCentroid
LeafNameCosineSimilarityToSearchQueriesCentroid
LeafNameJaccardSimilarityToWatchCentroid
LeafNameCosineSimilarityToWatchCentroid
LeafNameJaccardSimilarityToWatchCentroidBadge
LeafNameCosineSimilarityToWatchCentroidBadge
LeafNameJaccardSimilarityToTransactionCentroid
LeafNameCosineSimilarityToTransactionCentroid
LeafNameJaccardSimilarityToShoppingcartCentroid
LeafNameCosineSimilarityToShoppingcartCentroid

#L2 category name similarity features
L2NameJaccardSimilarityToRviCentroid
L2NameCosineSimilarityToRviCentroid
L2NameCosineSimilarityToFirstRvi
L2NameCosineSimilarityToFirstTwoRvisCentroid
L2NameCosineSimilarityToFirstThreeRvisCentroid
L2NameCosineSimilarityToFirstFiveRvisCentroid
L2NameCosineSimilarityCentroidRvisInLastDay
L2NameCosineSimilarityCentroidRvisInLastTwoDays
L2NameCosineSimilarityCentroidRvisInLastThreeDays
L2NameCosineSimilarityCentroidRvisInLastFiveDays
L2NameCosineSimilarityCentroidRvisInLastWeek
L2NameCosineSimilarityCentroidRvisInLastHour
L2NameCosineSimilarityCentroidRvisInLastTwoHours
L2NameCosineSimilarityCentroidRvisInLastFiveHours
L2NameCosineSimilarityCentroidRvisInLastHalfDay
L2NameCosineSimilarityWeightedCentroidRvisRankDist
L2NameCosineSimilarityWeightedCentroidRvisTimeDecay
L2NameJaccardSimilarityToSearchQueriesCentroid
L2NameCosineSimilarityToSearchQueriesCentroid
L2NameJaccardSimilarityToWatchCentroid
L2NameCosineSimilarityToWatchCentroid
L2NameJaccardSimilarityToWatchCentroidBadge
L2NameCosineSimilarityToWatchCentroidBadge
```

```
L2NameJaccardSimilarityToTransactionCentroid
L2NameCosineSimilarityToTransactionCentroid
L2NameJaccardSimilarityToShoppingcartCentroid
L2NameCosineSimilarityToShoppingcartCentroid

#Meta category name similarity features
MetaNameJaccardSimilarityToRviCentroid
MetaNameCosineSimilarityToRviCentroid
MetaNameCosineSimilarityToFirstRvi
MetaNameCosineSimilarityToFirstTwoRvisCentroid
MetaNameCosineSimilarityToFirstThreeRvisCentroid
MetaNameCosineSimilarityToFirstFiveRvisCentroid
MetaNameCosineSimilarityCentroidRvisInLastDay
MetaNameCosineSimilarityCentroidRvisInLastTwoDays
MetaNameCosineSimilarityCentroidRvisInLastThreeDays
MetaNameCosineSimilarityCentroidRvisInLastFiveDays
MetaNameCosineSimilarityCentroidRvisInLastWeek
MetaNameCosineSimilarityCentroidRvisInLastHour
MetaNameCosineSimilarityCentroidRvisInLastTwoHours
MetaNameCosineSimilarityCentroidRvisInLastFiveHours
MetaNameCosineSimilarityCentroidRvisInLastHalfDay
MetaNameCosineSimilarityWeightedCentroidRvisRankDist
MetaNameCosineSimilarityWeightedCentroidRvisTimeDecay
MetaNameJaccardSimilarityToSearchQueriesCentroid
MetaNameCosineSimilarityToSearchQueriesCentroid
MetaNameJaccardSimilarityToWatchCentroid
MetaNameCosineSimilarityToWatchCentroid
MetaNameJaccardSimilarityToWatchCentroidBadge
MetaNameCosineSimilarityToWatchCentroidBadge
MetaNameJaccardSimilarityToTransactionCentroid
MetaNameCosineSimilarityToTransactionCentroid
MetaNameJaccardSimilarityToShoppingcartCentroid
MetaNameCosineSimilarityToShoppingcartCentroid
```