Contributors: Wajih Arfaoui, Dilda Zhaksybek, and Dimitri Kestenbaum
IÉSEG School of Management MBD Big Data Analytics

Overview

The purpose of this report is to document and benchmark different classes of recommendation system algorithms our team applied with the objective of improving LastFMs current recommendation system (non-personalized top 10 recommender). Our ultimate goal with this project is to replace LastFMs simplistic recommender system with a more sophisticated model that provides diverse personalized recommendations that will significantly boost the quality of UX, subsequently improving LastFMs product. The recommendation techniques that will be explored and benchmarked include:

- Collaborative Filtering
- Content-Based Recommendation
- Hybrid Recommenders

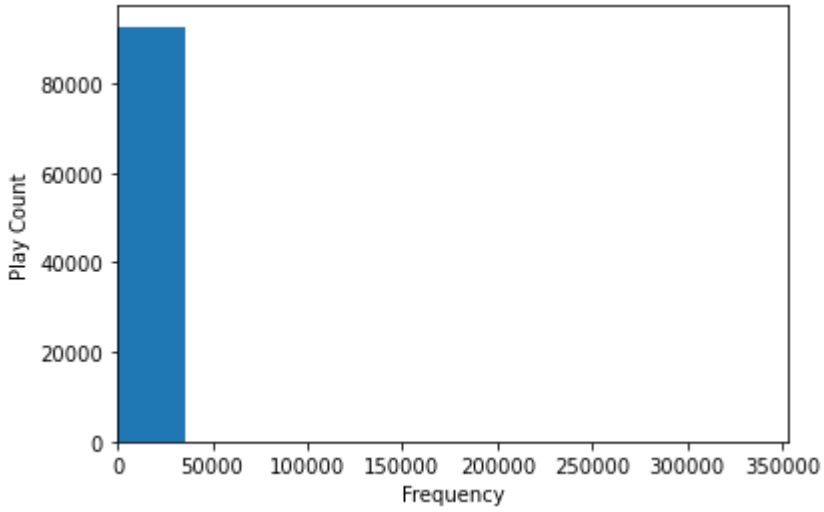
Collaborative Filtering Models Findings Report

Collaborative filtering hinges on the idea that we can predict a users preferences based on the preferences of users as well as predict the rating a user would giveitems based on their interactions with other items. It uses matrices of rating scores indexed by users and items respectively, and then computes a similarity measure: `cosine similarity` or `pearson correlation`, between vectors of ratings. The largest drawback to collaborative filtering models is that if they lack information for a new product or user, this poses the cold-start issue. This means predictions cannot be made as they would for users and items with more complete information on interaction.

Data Preprocessing for `Surprise`

In terms of data wrangling, there was a minimal amount of manipulation required in order to produce a dataframe which would be compatible with classes from the Surprise library. Surprise requires data to follow a three-column structure of user ID, item id (in this case `artistID`), and a rating. As the `user_artist.dat` file already complied with this requirement no columns were dropped.

However, instead of a discrete rating column, the `user_artist` data contains a continuous feature `weight` which indicates the number of songs played per artist for each unique user. A little visualization demonstrated a large amount of skew in this data:



We can see in the figure above how left-skewed the `weight` data is. However, due to the scale of the horizontal axis, which represents frequency, the right-tail distribution caused by the skew is obfuscated from the chart. In fact the maximum `weight` is actually `352698` plays, but only has a frequency of one, therefore does not appear in this plot.

Having a continuous measure like `weight` is not compatible with `Surprise`, which requires some kind of discrete rating scale such as 1-5. Therefore, the next step was too discretize or categorize the `weight` column to make it conform to a scale. A 1-5 rating scale was used as the distribution of `weight` could be equally divided in five quantiles. The code chunk below demonstrates how this discretization was achieved using the `qcut` function from `Pandas`.

```
#discretize weights using qcuts
user_artists_df['weight_quantiles'] = pd.qcut(user_artists_df['weight'],
                            q=[0,.2,.4,.6,.8,1],
                            labels=False,
                            precision=0)
```

Upon categorizing the `weight` and renaming it `weight_quantiles` the observations are almost equally distributed within these categories the exact values counts can be seen below:

- 0 18770
- 3 18581
- 4 18548
- 2 18469
- 1 18466

`Train`/`Test` Split + Initializing Surprise Objects

Next we perform a traditional train/test split of the `user_artists` data with a test-set size of `0.3`. Next is a crucial step when working with the `Surprise` library. The `Reader` class is used to parse our brand new ratings (`weight_quantiles`) based on given scale (1-5 of course). Secondly, a full dataset object is created for subsequent cross-validation as well as `Surprise`-specific trainset and testset objects. The code for this can be seen below:

```
#create reader object
reader = surprise.Reader(rating_scale=(1,5)) #1:5 scale

#create surprise train and test set objects
data =
surprise.Dataset.load_from_df(UA_df_cf[["userID","artistID","weight_quanti
les"]], reader)
UA_train = surprise.Dataset.load_from_df(UA_train,
reader).build_full_trainset()
UA_test = list(UA_test.itertuples(index=False, name=None))
```

Baseline Predictions

The next step is to explore what collaborative filtering algorithms or models from `Surprise` package give us a decent baseline or benchmark estimation by evaluating predictions based on the following metrics:

- `RMSE`
- `MAE`
- `NDCG`
- `F1-Score`

User-based KNNBasic:

```
#get baseline KNN score
from surprise import KNNBasic

# create options dict; use cosine similarity on user_based data
options = {'name':'cosine', 'user_based':True}

ubKNN = KNNBasic(k=20, min_k=2, sim_options=options, random_state=123)

#create cosine similarity matrix
ubKNN.fit(UA_train)\
.compute_similarities()
```

```
Impossible: 0.2671
```

Impossibility here indicates the porportion of the predictions that weren't possible due to cold-start issues (incomplete data). This is an important caveat considering this is one of the biggest weaknesses of collaborative filtering methods.

Baseline Results:

RMSE	1.422396
MAE	1.204500
Recall	0.002518
Precision	0.175000
F1	0.004965
NDCG@5	0.859066

Item-based KNNBasic:

```
#get baseline KNN score
from surprise import KNNBasic

# create options dict; use cosine similarity on user_based data
options = {'name':'cosine', 'user_based':False}

ibKNN = KNNBasic(k=20, min_k=2, sim_options=options, random_state=123)

#create cosine similarity matrix
ibKNN.fit(UA_train)\
.compute_similarities()
```

```
Impossible: 0.2012
```

Baseline Results:

RMSE	1.080457
MAE	0.866984
Recall	0.156475
Precision	0.915789
F1	0.267281
NDCG@5	0.859624

BaselineOnly (ALS)

```
from surprise import BaselineOnly

#alternating least squares (ALS) with 30 iterations
options = {"method": "als", "n_epochs": 30}
als = BaselineOnly(bsl_options=options)

# fit on training set
als.fit(UA_train)

als_preds = als.test(UA_test)
als_accuracy = accuracy.rmse(als_preds)
```

```
Impossible: 0.0000
```

Baseline Results:

RMSE	0.939793
MAE	0.786951
Recall	0.004856
Precision	0.870968
F1	0.009658
NDCG@5	0.864142

SVD (Matrix Factorization

```
from surprise import SVD

# 20 factors non-bias
svd = SVD(n_factors=20, biased=True, random_state=42)

# fit on training set
svd.fit(UA_train)

svd_preds = svd.test(UA_test)
svd_accuracy = accuracy.rmse(svd_preds)
```

```
Impossible: 0.0000
```

Baseline Results:

RMSE	0.906083
MAE	0.736482
Recall	0.063669
Precision	0.919481
F1	0.119092
NDCG@5	0.870266

CoClustering

```
from surprise import CoClustering

clust = CoClustering(n_cltr_u=10, n_cltr_i=10, n_epochs=50,
random_state=42)

clust.fit(UA_train)

coclust_preds = clust.test(UA_test)
accuracy.rmse(coclust_preds)
```

```
Impossible: 0.0000
```

Baseline Results:

RMSE	1.074960
MAE	0.864092
Recall	0.156115
Precision	0.745704
F1	0.258180
NDCG@5	0.860111

Baseline Consensus:

By comparing evaluation metrics we can see there's a clear winner in terms of baseline perfomance, `SVD`. Which trumps all other models across all evaluation metrics besides `recall`.

Hyperparameter Tuning

Given we've identified which model works best with our data. Next we perform a cross-validated grid search to find which hyperparameter combinations produce the best results. The grid search we ran can be seen in the code chunk below.

```
grid search cross validation

param_grid = {'n_factors':[20,50,150,200],
              #'n_epochs':[20,50],
              #'lr_all':[0.005, 0.001, 0.002],
              #'biased':[True],
              #'reg_all':[0.01,0.02]}

3-fold grid searched cv
SVD_cv_grid = GridSearchCV(SVD,
                           param_grid=param_grid,
                           measures=['rmse','mae'],
                           cv=3)
SVD_cv_grid.fit(data)
```

Final Model `train`, `predict`, and `evaluate`

```
# with tuned hyperparameters
svd_best = SVD(n_factors=20, n_epochs=20,lr_all=0.005, biased=True,
reg_all=0.01, random_state=42)

# fit on training set
svd_best.fit(UA_train)
```

RMSE	0.906055
MAE	0.735878
Recall	0.067806
Precision	0.915049
F1	0.126256
NDCG@5	0.870867

We can see from the above evaluation metrics we have indeed lowered the `RMSE` of the baseline predictions by a very small relatively negligible margin. Through this benchmarking evaluation process we have select the `SVD` model as the most effective collaborative filtering model for this problem.

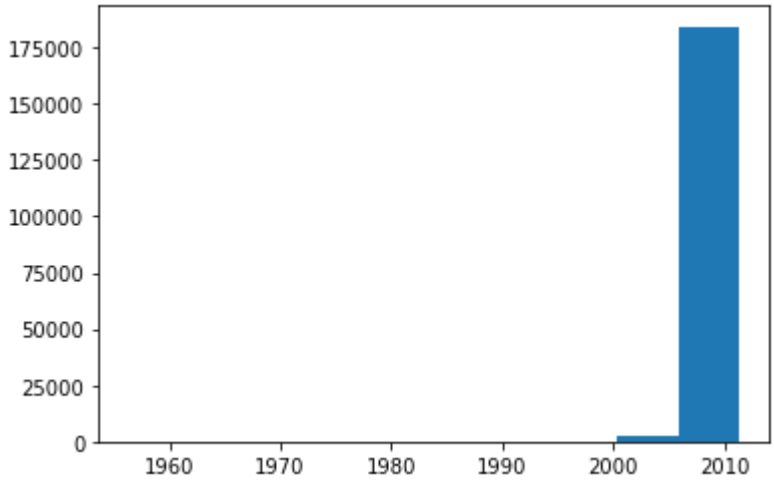
Content Based Recommendation System

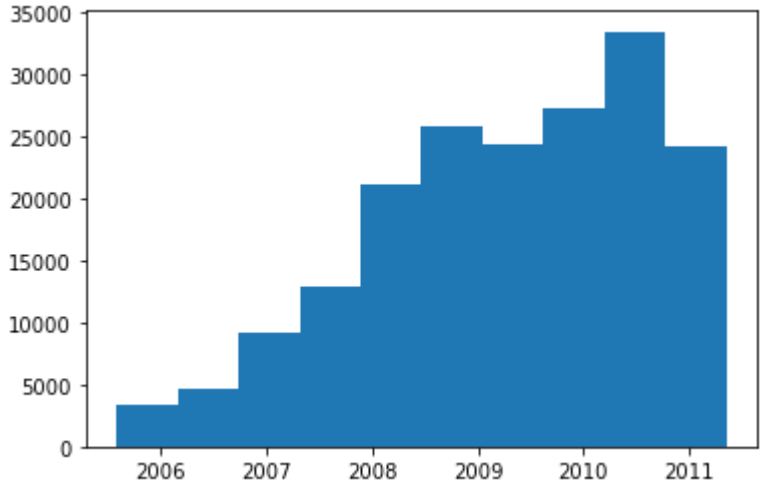
Content based systems are based on the description of the item and user's preference towards the item. In our case we create a content matrix of genres and recency for each artist, and fit on the ratings assigned by the users for these artists to come up with recommendations for the users. The system will recommend the artists similar to those user already likes based on the cosine similarities matrix. Content Based Recommendation Systems are considered very adaptive and user-tailored. They can make highly precise recommendation for each user. Since the similarity is computed between artists/items, new artists can be suggested regardless of how many users rated them. However, similar to collaborative filtering systems, content based recommender system gives a cold start for the new users, as there is not much data available to make a prediction. Besides, content based recommender systems tend to overspecialize, by recommending artists for the users that are very similar to those they already listen to, keeping the user in the same consumption bubble.

The matrix for Content based recommendation systems is created using two data frames: user_taggedartists.dat and tags.dat. We further discuss steps executed on these datasets in order to create the content matrix.

Data preprocessing and content matrix for content-based recommendation system

To being, we create a full-date variable column on user_taggedartists.dat. We can see that dates when users tagged artists are mostly frm 2000 and on.





We then proceed to create qualitative variables for the same data frame, and a recency variable column later, by categorizing dates artists were tagged by a user as "Very Old" if tagged before January 1970, "Old" if tagged before Jaunary 1984, "New" if tagged before January 2010 and "Very New" from January 2010 and further.

We then merge the two data frames, one containing a tag value of each tag ID, and the other one containg UserID, Artist ID and Recency value.

To create content matrix, we pivot the merged table twice in order to turn categorical variables in dummy variables. Then, two pivoted tables merged together are combined into a content matrix, with ArtistID as an index and Genre and Recency as variables. The resulting matrix is:

																tagValue					recency
tagValue	0	'80s	-pearl fashion music	0 play yet	00	00's	007	00s	00s rock	1	...	zombie rave	zombieland	zoocore	zornish	ztt	zu	new	old	very_new	very_old
artistID																					
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	42.0	0.0	3.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	240.0	0.0	84.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	24.0	0.0	3.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	36.0	0.0	8.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	8.0	0.0	6.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18741	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0
18742	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
18743	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
18744	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	4.0	0.0	10.0	0.0
18745	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Applying content based model

We initiate the model at NN = 10, filtering the matrix for 10 nearest neighbors with non-negative similarity. Then, we fit on content using content matrix, and on ratings using train dataset.

```
# init content-based
cb_mod = ContentBased(NN=10)

# fit on content
cb_mod.fit(cb)

# fit on train_ratings
cb_mod.fit_ratings(UA_train)

cb_pred = cb_mod.test(UA_test)
```

We get the following evaluation results:

```
# compute metrics for CB RS
cb_res = eval.evaluate(cb_pred, topn=5, rating_cutoff=3.5).rename(columns=
{'value':'Content_based_10'})
cb_res
```

	Content_based_10
RMSE	0.891378
MAE	0.669223
Recall	0.386871
Precision	0.760339
F1	0.512814
NDCG@5	0.872620

Comparing to collaborative filtering, so far, content based gives a better performance. But let's take a look at hybrid recommendation systems.

Hybrid Recommender Systems

Hybrid recommendation systems are not a fully independent class of algorithms. Rather they blend methods together in an effort to mitigate the drawbacks of individual techniques. In our case, we've experimented with a hybrid of our two best performing models: The SVD model fit with tuned hyperparameters and the content based model seen in the previous section. We also implemented a`Random Forest` hybrid model which is trained on the predictions of both the content based and SVD models and as well as the actual rating values as a target variable and then makes predictions of its own. However, since hybrid systems aggregate different techniques they can inherit the drawbacks of the models used as well.

SVD and Content Based Hybrid (Weighting)

We applied a weighting method to prioritize the ratings of the better-performing model (content based), when aggregating the predicted ratings. No additional preprocessing was required for this. The code chunk below demonstrates how the weighting technique takes `0.6` (60%) of a content based prediction and then adds the remaining `0.4` (40%) to create a new hybrid prediction, can be seen below:

```
#Combine predictions (weights)

#extract predictions content-based and item-based
df_pred_cb, df_pred_svd = pd.DataFrame(cb_pred), pd.DataFrame(svd_preds)

df_hybrid = df_pred_cb.copy()
df_hybrid['est'] = (np.array(df_pred_cb['est'])*0.6) +
(np.array(df_pred_svd['est'])*0.4)
```

Model Performance:

Below we can see that ultimately this results in a lower `RMSE` than either of the individuals models produced of `0.871708`. However, we see some trade off for some of the other metrics such as `MAE`,`Recall`, and `NDCG@5`, which did not perform better than the individual models with respect to each indivual metric that was just enumerated.

RMSE	0.871708
MAE	0.690477
Recall	0.496712
Precision	0.930947
F1	0.647791
NDCG@5	0.861735

Random Forest Hybrid

Data Preprocesing:

The first step here is creating a new train test split to evaluate this models performance. Next we create a new data frame that contains three columns. A target column with the actual labels for each oberservation as well as a column for the predictions of both the SVD model and the content based model respectively. Finally, these are split into a df `X` which contains the features, and a vector `y` which contains the labels. The code for this can be seen in the code chunk below:

```
#Random Forest Hybrid
df_cb_train, df_cb_test = df_pred_cb[:19500], df_pred_cb[19500:]
df_svd_train, df_svd_test = df_pred_svd[:19500], df_pred_svd[19500:]

rf_data = df_cb_train[['r_ui', 'est']].rename(columns={'r_ui':'target',
'est':'cb_pred'})
rf_data['svd_pred'] = df_svd_train['est']

rf_test = df_cb_test[['r_ui', 'est']].rename(columns={'r_ui':'target',
'est':'cb_pred'})
rf_test['svd_pred'] = df_svd_test['est']

from sklearn.ensemble import RandomForestRegressor

X, y = rf_data.loc[:,rf_data.columns != 'target'],
np.array(rf_data['target'])
X_test = rf_test.loc[:,rf_test.columns != 'target']
```

```
#fit random forest model
rf_model = RandomForestRegressor(max_depth=4, n_estimators=100).fit(X,y)

#predict
rf_pred = rf_model.predict(X_test)

# transform in surprise format
df_rf = df_cb_test.copy()
df_rf['est'] = rf_pred
```

Conclusion

Our Proposal

After applying different recommendation system models to the LastFM datasets at our disposal to compare and evaluate them, we've developed a comprehensive recommendation system strategy that we believe should replace LastFM's current non-personalized approach. This new system would rely directly on a mixture of non-personalized, collaborative filtering, and hybrid recommender systems. We believe a hybrid model is the most effective tool for the job in terms of percision since it takes into account the artist tags and similaritites between artists and users. Hence, it could mitigate the pitfalls common to recommendation systems such as overspecialization in content-based approaches (not providing a broad enough range of options) and cold-start issues related to collaborative filtering techniques, to ultimately provide novel and accurate predictions for different kinds of users.

Why Recommendation Systems Shouldn't Be One-Size-Fits-All

As an enhancement to LastFM's non-personalized recommendation system, and with the ubiquity of data collection, we saw that different recommandation strategies can be applied differently based on the context of each experience, such as the webpage to deploy it on, as well as the user's recency.

To do so, we thought about categorizing our users into 3 different types:

- First-time visitor: for this type of users, that we don't have any historical data of their musical taste and preferences, we can display `Top artists` on the homepage. The system will give more weight to recent ratings over historical ones and updates scores every time a data feed is synchronized (*non-personalised method*).
- Seasonal visitor: for users that don't have frequent activities on the platform, we may base our recommendations to them on a more user-specific strategy which is `Most Popular in Category` strategy, that not only promote the most popular artists but includes artist from music genres that they already liked and music genres that are similar to them (*item-based method*).
- Frequent visitor: this category, will have the `Most Personnalised Recommendation` since we already collected an important amount of data about their behaviours. The system that we can deploy will surface suggestions of artists for our users based on the user's personal similarity with other users, the similarities of artists in addition to the genres and release date affinity (*hybrid method*).