MIT Capstone Project

Recommendation systems by Karan Uppal

What is the problem to be solved?

I. Creating a best fit model using recommendation systems to extract the desired recommended songs from a large dataset.

II. Noting that play counts and the type of songs played within specific user profiles inherently provide important information of what songs are popular versus which are not.

What are key patterns within data that we must account for?

- Pre-processing of 21st century songs vs 20th century songs.

* This was accounted for using User ID information to notice how popular the song was within each individuals playlist to account for varied popularity within different time periods.

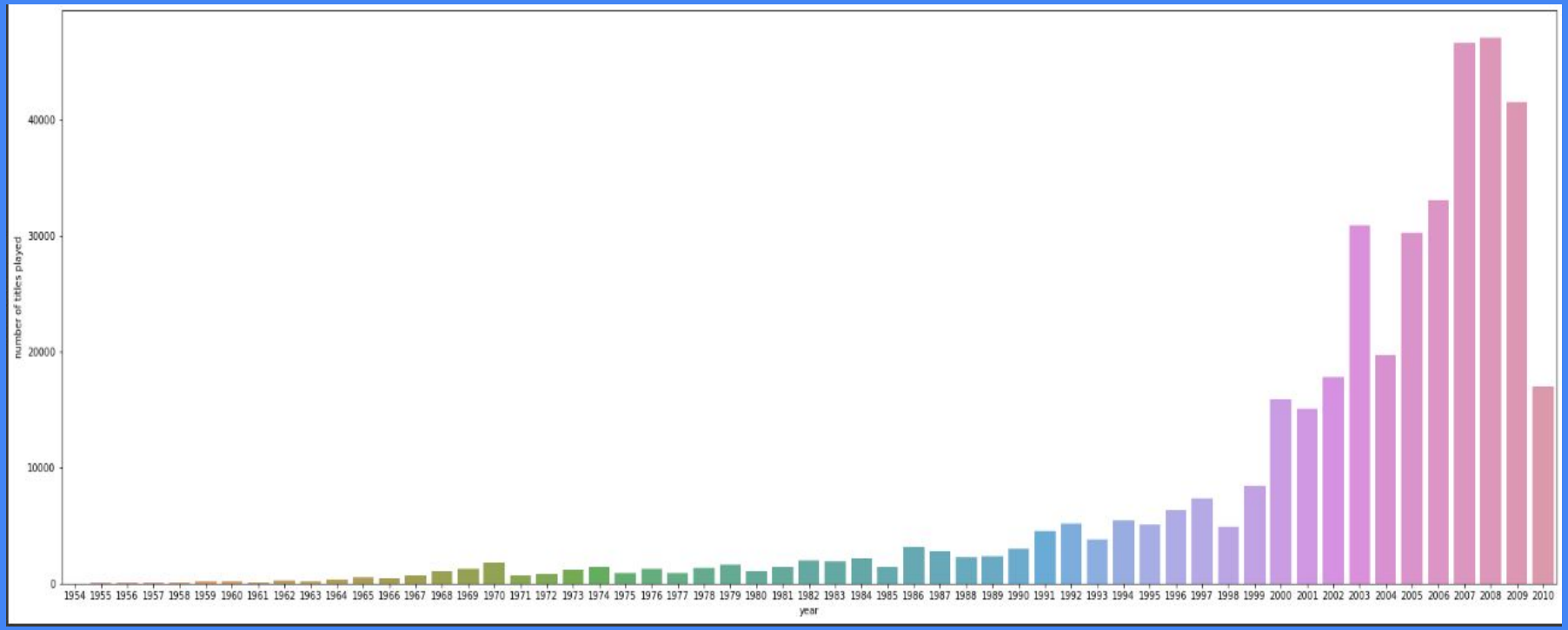

Figure 1: Play count of different songs based on release dates

*Notice the distribution of songs played in the 21st century versus the 20th century. This is most likely due to the technological age difference and exposure of music to the public.

Techniques used creating the recommendation system

- Marking statistical relationships between different recommendation systems (i.e RMSE, Recall, Precision, and F_1 score)
- Improving RMSE score based on baseline model by directly increasing the relevancy of the recommendation system
- Prioritizing recall and its significance based on dataset
- Using SVD to predict how the users would rate the items so the users can get the recommendation based on the prediction since release of songs are dispersed.(Find latent features between songs and users)

Insight provided from Recommendation systems

	User-User Similarity-Based Collaborative Filtering	Item-item Similarity-based collaborative filtering	Matrix Factorization	Cluster **No statistical significance*
RMSE	4.7030 => 4.6693	4.6135 => 4.6391	4.6031 => 5.3422	4.5942 => 4.5941
Precision	0.447 => 0.487	0.396 => 0.434	0.489 => 0.433	0.452 => 0.453
Recall	0.851 => 0.781	0.694 => 0.645	0.77 => 0.915	0.641 => 0.642
F_1 Score	0.586 => 0.6	0.504 => 0.519	0.598 => 0.588	0.53 => 0.531

Proposed Solution Design

I. Matrix Factorization optimization

II. It's latent features with regards to 21st century and 20th century songs were balanced based on individual profiles

III. High recall rate (92%) and statistically unchanged F_1 score

Why is this method valid compared to other models?

- RMSE and Recall are improved by over 16% and 19% respectively. This means the overall predicted ratings are very similar to the actual ratings

- Matrix Factorization provides a F_1 score and Precision that were not changed by any significant margin compared to other models.

- Most recommendations that were relevant (F_1 score for Matrix Factorization was already high relative to all other models)

What are the key actionables for stakeholders?

- Model accounts for the time the song was released with regards to what user finds popular and excludes what is not

- Using this system provides relevant movies to users based on their historical interactions simply only improves customer satisfaction and hence improve revenue since system accommodates to user profile.

- Model provides relevant recommendations and although precision and ts F_1 score were lowered, in comparison to other models they still are not statistically insignificant.

- The suggested model is based on the past behavior of the user and it is not dependent on any additional information tracking the behavior and similarities of the data to the predictions/recommendations.

Further improvements to recommended model

1. Multiple users on a single profile while using this system may disrupt the consistency of behavior
   - ➢ Create multiple profiles on a single user as used with Netflix
2. Changes in the behavior of how people make song selections may affect profiling
   - ➢ Track a majority model of behavior and see if irregular behaviors align with primary model to eliminate outlier behavior while choosing songs
3. Since matrix factorization uses relationships of latent features with users and items, some associations might not be affiliated with that individual and eliminating those associations might be difficult while making recommendations
   - ➢ Use of a sparse matrix might be needed to dissociate these findings as the use of a content based filtering system might also be necessary to take into account of extracted features from the text data. This helped find similarities between songs using these features using cosine function.